Genotype-Phenotype Maps in Complex Living Systems

Dissertation

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Jose Aguilar Rodriguez
aus
Spanien

Promotionskommission
Prof. Dr. Andreas Wagner (Vorsitz)
Prof. Dr. Martin Ackermann
Prof. Dr. Owen Petchey

Zürich, 2017

To Marta, for everything.

To my family, for their support.

Almost in the beginning was curiosity.

ISAAC ASIMOV, Asimov's New Guide to Science

All men, by nature, desire to know.

ARISTOTLE, Metaphysics

To strive, to seek, to find, and not to yield.

ALFRED, LORD TENNYSON, Ulysses

En el sueño del hombre que soñaba, el soñado se despertó.

JORGE LUIS BORGES, Las ruinas circulares

List of papers

This thesis is based on the following papers:

I. Chapter 2: *__Aguilar-Rodríguez, J.__, *Payne, J. L., Wagner, A. (2017). "A thousand empirical adaptive landscapes and their navigability." *Nature Ecology & Evolution*, 1: 0045.

II. Chapter 3: **Aguilar-Rodríguez, J.**, Peel, L., Stella, M., †Wagner, A., †Payne, J. L. "The architecture of an empirical genotype-phenotype map." *Submitted.*

III. Chapter 4: *__Aguilar-Rodríguez, J.__, *Sabater-Muñoz, B., Montagud-Martínez, R., Berlanga, V., Alvarez-Ponce, D., †Wagner, A., †Fares, M. A. (2016). "The molecular chaperone DnaK is a source of mutational robustness." *Genome Biology and Evolution*, 8: 2979–2991.

IV. Chapter 5: **Aguilar-Rodríguez, J.**, Fares, M. A., Wagner, A. "Chaperonin overproduction and metabolic erosion caused by mutation accumulation in *Escherichia coli*." *Submitted.*

V. Chapter 6: **Aguilar-Rodríguez, J.**, Wagner, A. "Metabolic determinants of enzyme evolution in a genome-scale bacterial metabolic network." *Submitted.*

*Equal contribution

†Co-corresponding authors

Contents

Summary

The mapping of genotypes onto phenotypes is one of the most fundamental endeavors in biology, with important consequences for evolution, development, and disease. Most of what we know about genotype-phenotype maps comes from ever more sophisticated computational models of biological systems. However, the study of genotype-phenotype maps is currently shifting away from the theoretical models that shaped the field, toward experimental data derived from high-throughput technologies. In this thesis, I contribute to this shift by embracing a systems-level and evolutionary perspective to study genotype-phenotype maps of different complex biological systems at multiple levels of biological organization.

In chapter 2, I exhaustively analyze 1,137 empirical and complete genotype-phenotype landscapes, each describing the binding affinity of a eukaryotic transcription factor to all possible short DNA sequences. I find that these landscapes are highly navigable through single mutations and natural selection, indicating that the regulatory effect of binding is readily fine-tuned via mutations in transcription factor binding sites. These landscapes have few peaks that comprise dozens to hundreds of sequences, and that vary in their evolutionary accessibility. These findings, which are based on *in vitro* data, are supported by three additional analyses that are based on *in vivo* data. First, in *Mus musculus*, high-affinity transcription factor binding sites from rugged landscapes are less prevalent in protein-bound regions of the genome than high-affinity sites from smooth landscapes. Second, in *Saccharomyces cerevisiae*, gene expression measurements from hundreds of engineered promoters closely reflect landscape topography. And third, the amount of genetic polymorphism in binding sites in *S. cerevisiae* increases with the number of sequences in a peak. Together, these analyses indicate that landscape topography has

helped shape the portfolio of regulatory DNA in two highly diverged eukaryotic species, and may have contributed to the enormous success of transcriptional regulation as a source of evolutionary novelties.

In chapter 3, I study an empirical genotype-phenotype map of transcription factor binding preferences. In this map, genotypes are short DNA sequences and phenotypes are the transcription factors that bind these sequences. I study the internal structure of networks describing the mutational connections between genotypes mapping onto the same phenotype, and how these genotype networks interface and overlap with one another in the space of all possible binding sites. In so doing, I provide a high-resolution depiction of the architecture of an empirical genotype-phenotype map. I show that these genotype networks are assortative, "small-world," and tend to overlap and interface with one another. I discuss the implications that these findings have for the evolution of gene regulation.

In chapters 4 and 5, I study how molecular chaperones alter the mapping from the genotypes to the phenotypes of proteins, and the evolutionary consequences that these modified protein genotype-phenotype maps have on genome evolution. In chapter 4, I analyze evolutionary rates of proteins that require the bacterial chaperone DnaK for folding through a combination of experimental and comparative approaches. Most of the evidence I find indicates that DnaK can buffer deleterious mutations in its target proteins, and that these proteins therefore evolve faster than in the absence of DnaK-mediated folding. This is the first demonstration that a member of the Hsp70 family of chaperones can buffer the effect of destabilizing mutations, with long-term consequences on protein evolution. In chapter 5, I study how the metabolic erosion experienced by hypermutable populations of *Escherichia coli* evolved for thousands of generations in a mutation accumulation experiment can be ameliorated in some environments thanks to the mutational buffering provided by the bacterial chaperonin GroEL.

Finally, in chapter 6, I analyze the mapping from metabolic genotypes—a genome's set of enzyme-encoding genes—to metabolic phenotypes—the set of molecules a metabolism

can synthesize. Specifically, I study how selection for a given metabolic phenotype can constrain enzyme evolution in the genome-scale metabolic network of *E. coli*. Central and highly connected enzymes do not evolve more slowly than less connected enzymes because of their position in the metabolic network. In contrast, enzymes catalyzing reactions with high metabolic flux—high substrate to product conversion rates—evolve slowly. Moreover, enzymes catalyzing reactions that are essential in many different genetic backgrounds also evolve more slowly. My analyses show that an enzyme's contribution to the function of a metabolic network affects its evolution more than its location in the network's structure.

Zusammenfassung

Die Zuordnung von Genotypen zu Phänotypen ist ein fundamentales Ziel biologischer Forschung, mit wichtigen Auswirkungen auf das Verständnis der Evolution, der Embryonalentwicklung und des Entstehens von Krankheiten. Bisherige Erkenntnisse aus Genotyp-Phänotyp Karten («genotype-phenotype maps») stammen von ausgeklügelten Computermodellen, die auf biologischen Systemen basieren. Die zeitgenössische Forschung bewegt sich allerdings weg von den theoretischen Modellen, die das Feld bisher geprägt haben, und nutzt vermehrt Daten, die aus «high-throughput» Experimenten gewonnen wurden, um Genotyp-Phänotyp Karten zu generieren. Mit meiner Doktorarbeit trage ich zu diesem Trend bei, indem ich einen Ansatz wählte, der eine evolutionäre und systembiologische Perspektive nutzt, um Genotyp-Phänotyp Karten von komplexen Systemen auf mehreren organisatorischen Ebenen zu untersuchen.

In Kapitel 2 analysiere ich 1137 empirische und vollständige Genotyp-Phänotyp Landschaften, welche die Bindungsaffinität eines eukaryotischen Transkriptionsfaktors an alle möglichen kurzen DNA Sequenzen beschreiben. Ich habe herausgefunden, dass diese Landschaften durch Punktmutationen und natürliche Auslese gut navigierbar sein können, was darauf hinweist, dass der regulatorische Effekt der Bindung durch Mutationen in der Bindungsstelle eines Transkriptionsfaktors fein abgestimmt sein kann. Diese Landschaften haben wenige «Gipfel», die aus dutzenden bis hunderten kurzen Sequenzen bestehen können und sich in ihrer Zugänglichkeit Unterscheiden. Dieses Ergebnis, welches auf *in vitro* Daten basiert, wird zusätzlich noch von drei weiteren *in vivo* Studien unterstützt. Die Erste dieser Studien, in *Mus musculus*, zeigt, dass Hochaffinitäts-Bindungsregionen von Transkriptionsfaktoren in zerklüfteten («rugged») Landschaften sich weniger häufig in proteingebundenen Regionen des Genoms befinden als Hochaffinitäts-Regionen von

glatten («smooth») Landschaften. Die zweite Studie, in *Saccharomyces cerevisiae*, zeigt, dass die Expression von hunderten modifizierten Promotoren eine genaue Landschaftstopographie wiederspiegelt. In der dritten Studie wird deutlich, dass in *S. cerevisiae* die Anzahl von genetischen Polymorphismen in Bindungsstellen mit der Anzahl der Sequenzen in einem Landschaftsgipfel steigt. Diese Analyse deutet darauf hin, dass die Landschaftstopographie geholfen hat, den Bestand der regulatorischen DNA in zwei sehr unterschiedlichen eukariotischen Arten zu formen und wahrscheinlich zum enormen Erfolg der transkriptionellen Regulation als Quelle für evolutionäre Innovation beitrug.

In Kapitel 3 untersuche ich eine empirische Genotyp-Phänotyp Karte auf Bindungspräferenzen von Transkriptionsfaktoren. In dieser Karte sind die Genotypen kurze DNA Sequenzen und die Phänotypen die Transkriptionsfaktoren, welche an diese kurzen DNA Sequenzen binden. Ich analysiere die interne Struktur von Genotypnetzwerken und beschreibe wie Mutationen in verschiedenen Genotypen den gleichen Phänotyp erzeugen können, und wie verschiedene Genotypnetzwerke überlappen. Dieser Ansatz bot mir die Möglichkeit eine hochauflösende, empirische Genotyp-Phänotyp Karte zu erstellen. Damit konnte ich zeigen, dass diese Genotypnetzwerke zusammenpassen und die Tendenz haben sich zu überlappen oder miteinander verbunden zu sein. Eingehend diskutiere ich die Auswirkungen dieser Forschungsergebnisse auf die Evolution der Genregulation.

In den Kapiteln 4 und 5 beschreibe ich, wie «Chaperone» die Zuordnung von Genotypen zu Phänotypen der Proteine verändern und wie sich diese modifizierten Protein-Genotyp-Phänotyp Karten auf die Genomevolution auswirken. Kapitel 4 beschreibt, wie ich eine Kombination aus experimentellen und vergleichenden Ansätzen benutze, um die evolutionären Raten von Proteinen, welche das bakterielle Chaperon DnaK für die Bildung ihrer korrekten Struktur benötigen, zu berechnen. Meine Resultate deuten darauf hin, dass DnaK die Auswirkung von schädlichen Mutationen in seinen Zielproteinen reduzieren kann und dass Proteine in der Abwesenheit von DnaK schneller evolvieren. Das ist das erste mal, dass gezeigt werden konnte, dass ein Protein aus der Hsp70 Familie der Chaperone schädliche Mutationen in seinen Zielproteinen reduzieren kann, mit direkten

Auswirkungen auf die Evolution dieser Proteine. In Kapitel 5 zeige ich wie metabolische Erosion, verursacht durch hypermutierende Populationen von *Escherichia coli*, welche für tausende von Generationen in einem Mutationsakkumulationsexperiment evolviert wurden, durch das bakterielle Chaperon GroEL abgeschwächt werden kann.

Im letzten Kapitel analysiere ich die Zuordnung von metabolischen Genotypen (die enzymkodierenden Gene eines Genoms) zu metabolischen Phänotypen (die Bandbreite von Molekülen, die ein Metabolismus synthetisieren kann). Ich untersuche, wie die Selektion für einen bestimmten metabolischen Phänotyp die Enzymevolution im metabolischen Netzwerk von *E. coli* hemmen kann. Zentrale und hochvernezte Enzyme evolvieren nicht langsamer als weniger vernetzte Enzyme. Hingegen evolvieren Enzyme mit einem hohen metabolischen Fluss, d.h. mit einer hohen Substratumsatzrate, langsam. Ausserdem evolvieren Enzyme langsamer, wenn sie an Reaktionen beteiligt sind, die in vielen genetischen Hintergründen essentiall sind. Meine Analyse zeigt, dass der Beitrag eines Enzyms zur Funktion eines metabolischen Netzwerks einen grösseren Effekt als seine Position in der Netzwerkstruktur auf die Evolution des Enzyms aufweist.

1 Introduction

It was the botanist and geneticist Wilhelm Johannsen who coined the terms "genotype" and "phenotype" in 1909 [1, 2], just a few years after introducing the word "gene" [3]. In current usage, genotype refers to the DNA sequence of an individual, while phenotype refers to the visible or measurable traits of an individual [4]. Phenotypes are the physical manifestations of genotypes. The relationship between genotypes and phenotypes is described by genotype-phenotype maps, which assign a phenotype to every possible genotype. Since the birth of genetics a major goal of biology has been the quantitative description of these maps in diverse biological systems.

The distinction between genotypes and phenotypes provided a useful framework to early geneticists, who like Mendel, could only infer genotypes from the inheritance patterns of phenotypes [4, 5]. The distinction also proved valuable for the study of evolution [6]. Mutations occur at the level of the genotype and generate heritable variation. This variation translates into phenotypic variation, which is the substrate of natural selection. Fisher, Haldane, and Wright—the founding fathers of population genetics [7]—assumed a simple mapping from genotypes onto phenotypes. This approach, inherited by all subsequent population geneticists, proved very successful to study the evolutionary dynamics of genotypes in a population. Population genetics focuses on the statistical effects of genes on phenotypes, and ignores the mechanistic understanding of biological systems, which is the focus of biochemistry and molecular biology [8–11]. The enormous success of population genetics is in part due to this disregard towards the complexity of genotype-phenotype maps. However, this disregard comes at a price, since the simple models of population genetics fail to capture many important evolutionary phenomena, such as evolutionary innovations [12, 13], or evolutionary constraints [14, 15].

Nowadays, evolutionary systems biology is a growing field that tries to remedy this caveat of common evolutionary thinking by integrating a mechanistic understanding of living systems into an evolutionary framework [16–24]. The main research goal of this young discipline is to study genotype-phenotype maps across different levels of biological organization. The study of genotype-phenotype maps is currently shifting away from the conceptual and computational models that shaped the field, toward empirical data derived from high-throughput technologies [25–34]. In this thesis, I contribute to this shift by embracing a systems-level and evolutionary perspective to study genotype-phenotype maps of different complex biological systems.

1.1 Genotype-phenotype maps

The mapping of genotypes onto phenotypes is one of the central undertakings in biology, with important implications for the study of evolution, development, and disease [24, 35–37]. Genotype-phenotype maps can be conceptualized for different biological systems at distinct levels of biological organization [12]. Here, I refer to a system as "a set of elements or parts that cooperate to perform a task" [12]. For instance, a protein enzyme is a system whose parts—amino acids—cooperate to catalyze a metabolic reaction. The genotype of a protein is its primary sequence of amino acids, while we may think of its phenotype as its three-dimensional tertiary structure, which is responsible for its biological function—catalysis in the case of an enzyme. Similarly, the genotype-phenotype concept can be applied to higher levels of biological organization, such as metabolism. A metabolic genotype is the set of enzymes present in a genome, while a metabolic phenotype is the set of molecules that these enzymes synthesize from nutrients. Genotype-phenotype maps can also be used to study man-made systems such as reconfigurable electronic circuitry [38], and digital organisms [39]. Therefore, genotype-phenotype maps are useful to study the evolution of a wide variety of systems, including both natural and technological systems.

As we will see, the idea behind genotype-phenotype maps can be traced back to the work of Sewall Wright [40] and John Maynard Smith [41]. However, the term genotype-phenotype map ("genotype-phenotype mapping") itself was only coined in 1970 by Jim

Burns [42], who outlined the research programme of evolutionary systems biology before the development of systems biology made it feasible [24]:

> It is the quantitative phenotype, arising from the genotypic prescriptions and the environment, which is of critical importance for the cell's survival and which therefore features in population genetic theory. A study of this synthetic problem would thus, by providing genotype-phenotype mappings for simple synthetic systems, help to connect two major areas of biological theory: the biochemical and the population genetic.

In particular, Burns was interested in genotype-phenotype maps in the context of cellular metabolism, a topic which I will develop in section 1.4. The term "genotype-phenotype map" was re-introduced in 1991 by the developmental biologist Pere Alberch as a useful concept for the integration of genetics into the study of the complex developmental processes that generate morphological phenotypes such as the vertebrate limb [35].

The genotype-fitness map is a particularly important type of genotype-phenotype map, especially for evolutionary research. To an evolutionary geneticist, fitness—the contribution of a specific genotype to future generations due to reproduction, differential survival or both [43–45]—is the ultimate phenotype, as it is the one upon which natural selection acts [10]. Because Sewall Wright envisioned such a map as a landscape where fitness defines the "elevation" of each coordinate in genotype space—the space of all possible genotypes—genotype-fitness maps are better known as adaptive (or fitness) landscapes [40]. Since the concept was introduced in the early 1930s [40, 46], theoretical fitness landscapes have received considerable attention from evolutionary biologists interested in understanding how landscape topography affects evolutionary dynamics. The studies resulting from this interest have shown that the topography of a fitness landscape has important evolutionary consequences, and specifically for speciation [47], the evolution of sex [48, 49], mutational robustness [50, 51] and the predictability of evolution [52–54], [55–58]. Given the tremendous importance of this particular type of map, I will explore it in greater detail in section 1.1.3.

Most of what we know about genotype-phenotype maps comes from computational models that predict phenotypes from genotypes in diverse biological systems [59–62]. Such models can be very sophisticated [63]; as, for instance, a whole-cell model of the life cycle of *Mycoplasma genitalium* [64], or a complex computational model for tooth development in mammals [65, 66]. However, most of our thinking about genotype-phenotype maps has been shaped by models that map RNA sequences onto secondary structures [60], binary amino acid sequences onto lattice-based structures [59], regulatory circuit genotypes onto gene expression patterns [61], as well as metabolic genotypes onto metabolite utilization phenotypes [62]. These computational models allow us to rapidly map genotypes to phenotypes in a comprehensive manner. Thanks to these models, it has been possible to study genotype-phenotype maps of molecular systems as diverse as RNA, proteins, regulatory networks, and metabolism (for an extensive review, see [12]). Despite the great differences between all of these biological systems, these theoretical studies have revealed some striking commonalities in their genotype-phenotype maps. First, epistasis—non-additive interaction between individual mutations—is pervasive (section 1.1.2). Second, these systems are to some extent robust to genotypic change. This robustness leads to the existence of genotype networks (aka neutral networks) (section 1.1.4).

The structure of a genotype-phenotype map has important evolutionary consequences. It can influence the accumulation of genetic diversity [50], the rate of adaptation [67, 68], the robustness and evolvability of genotypes and phenotypes [69], as well as their "findability" [70–72]. Therefore, it is highly important to move beyond theoretical models of genotype-phenotype maps and study maps derived from experimental data [37]. In recent years, the study of empirical genotype-phenotype maps and adaptive landscapes constructed from experimental data has become a burgeoning area of research [26, 29–32, 56, 73, 74]. This has been possible thanks to recent technological developments in high-throughput sequencing, DNA synthesis and lineage tracking, which are revolutionizing the experimental study of evolutionary processes [56, 75, 76]. The integration of these technologies has made it possible for the first time to assign phenotypic or fitness

values to a large number of genotypes. Now we can generate mutations and study their phenotypic or fitness effects precisely, cheaply, and in a highly parallel fashion. As a consequence of this, the characterization of empirical genotype-phenotype maps and real fitness landscapes is now much easier than ever before.

1.1.1 A conceptual classification of genotype-phenotype maps

A genotype-phenotype map is defined as the mathematical function that maps a set of genotypes into a set of phenotypes. These phenotypes can be either categorical or quantitative. A categorical phenotype is "a discrete classification that is assigned to each genotype" [77]. For example, the secondary structure of an RNA sequence is a categorical phenotype. As we shall see in chapter 3, in some cases, an individual genotype can have more than one categorical phenotype, such as an RNA sequence genotype that folds into multiple secondary structure phenotypes [78]. A quantitative phenotype is a real-valued phenotypic trait that can be assigned to each genotype. For example, we may think of the folding energy of an RNA sequence as a quantitative phenotype. Fig 1.1 shows a classificatory scheme for genotype-phenotype maps based on this distinction. Genotype-phenotypes landscapes are the subclass of genotype-phenotype maps where the phenotype is quantitative instead of categorical [30, 79]. Adaptive landscapes are an important subclass of genotype-phenotype landscapes, where fitness is the quantitative phenotype in consideration.

1.1.2 Epistasis

The term "epistasis" comprises all deviations from independent (additive) contributions of alleles at different loci on a particular phenotype [80, 81]. Epistasis means that the phenotypic effect of a mutation depends on the genetic background in which it emerges [82, 83]. Thus, epistasis can impose severe constraints on molecular evolution, because the substitutions that are beneficial in one background can be deleterious in another. Epistasis has been frequently demonstrated in RNA [84, 85], proteins [52, 86–90], metabolic networks [91–93], and gene regulatory circuits [94].

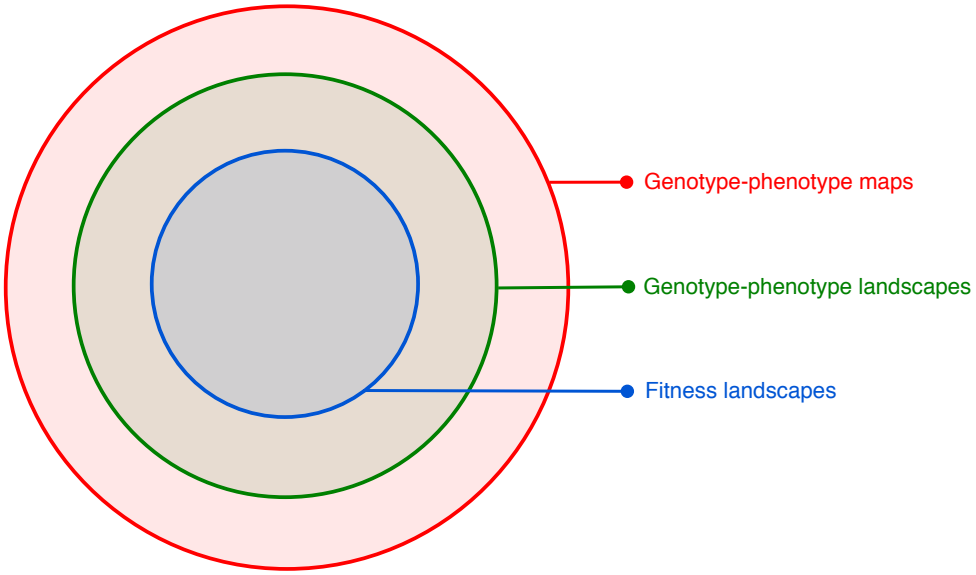


Figure 1.1: Classification scheme for genotype-phenotype maps. Genotype-phenotype maps describe the mapping from genotypes into phenotypes in a given biological system. Phenotypes can be either categorical or quantitative. Genotype-phenotype landscapes are a subclass of genotype-phenotype maps for which the phenotype is quantitative. Adaptive or fitness landscapes are a subclass of genotype-phenotype landscapes for which fitness is the quantitative phenotype. The relative sizes of the circles in this figure are not proportional to the relative importance of their associated type of genotype-phenotype map, either in nature or in the published literature on the topic.

The notion of epistasis is tightly related to the central question of how genotypes map onto phenotypes [21, 24]. Epistasis is important to understand the physical, biochemical and physiological basis of genotype-phenotype maps [80, 83], but also to comprehend their evolutionary causes and consequences [81]. The effect of epistasis on the architecture of these maps can be rapidly visualized with the powerful metaphor of the adaptive landscape [40, 95]. Graphical representations of an adaptive landscape often illustrate a genotype-phenotype (or genotype-fitness) landscape as a surface above a two-dimensional base plane representing genotype space, the space of all possible genotypes [40, 95]. The "height" of the landscape surface is determined by a real-valued measure of a phenotype (e.g. the folding energy of an RNA sequence). As I noted above, when the phenotype is biological fitness, then the landscape is called an adaptive landscape. Within this formulation, adaptation can be viewed as a hill-climbing process, where populations tend to move towards peaks as a consequence of mutation and natural selection. Epistasis

determines whether an adaptive landscape is smooth or rugged, that is, its topography [55]. For example, a specific type of epistasis, reciprocal sign epistasis, is a necessary condition for the presence of multiple peaks separated by low-fitness valleys [82]. This rugged topography can block the approach to the highest peak by causing entrapment on local suboptimal peaks [96]. In the absence of epistasis, adaptive landscapes are smooth and single peaked, and thus do not pose any obstacle to evolutionary exploration.

1.1.3 Adaptive landscapes

1.1.3.1 Historical background

The concept of the adaptive landscape was introduced in 1932 by Sewall Wright in a paper presented at the Sixth International Congress of Genetics in Ithaca, New York [40]. He was invited by E. M. East, together with R. A. Fisher and J. B. S. Haldane, to a session on the nascent discipline of population genetics [7]. East asked the three fathers of the discipline to make accessible their challenging mathematical work to an audience of general biologists with an interest in evolution, but easily intimidated by mathematics [97–99]. Wright's main work on populations genetics was his 159-page paper "Evolution in mendelian populations," published one year earlier [98]. The paper presented by Wright at the International Congress in 1932 was published in the proceedings under the title "The roles of mutation, inbreeding, crossbreeding and selection in evolution" [40]. It was in this less technical condensation of the 1931 paper that one of the most famous metaphors in the history of biology was publicly presented. The presence of diagrams of adaptive landscapes in all major evolution textbooks since 1937 testifies to the concept's enormous influence in evolutionary biology [100–104].

An adaptive landscape is commonly visualized as a "hilly" surface with peaks and valleys in a three-dimensional space. Each coordinate in the x–y plane represents an individual genotype. The third dimension (z axis), the "elevation" of the landscape, represents fitness. The regions of high elevation (high fitness) are called adaptive peaks, while the regions of low elevation (low fitness) are called adaptive valleys. However, Wright did not

use this popular representation and instead in 1932 employed two different representations (Fig. 1.2). The first of them represents the space of all possible genotypes, that is, genotype space, as a network of mutational relationships between different genotypes. In such a network, nodes represent genotypes, and two nodes are connected by an edge if their associated genotypes differ just by a single genetic change. Of course, Wright was aware that the genotype space for any organism must be incredibly large and high-dimensional, and estimated that its size would be on the order of 10^{1000}, and that it would have 1000 dimensions (one for each locus): "with 10 allelomorphs in each of 1000 loci, the number of possible combinations is 10^{1000} which is a very large number" [40]. It is worth noting that this large number is obtained trough a gross underestimation of the real number of genes and alleles in an average species. The second representation of an adaptive landscape used by Wright in 1932, and perhaps his most famous, is a two-dimensional graphical depiction akin to a topographic map (Fig. 1.2). It represents multidimensional genotype space as a two-dimensional plane, and the differences in fitness between different locations in the plane are visualized with the help of contour lines.

Wright's depictions of an adaptive landscape as a high-dimensional discrete genotype space was correctly identified as an hypercube by Haldane, who published it in print before Wright's 1932 paper [105]. However, Wright's correspondence shows that he thought about the adaptive landscape before Haldane [46, 104]. He already described the concept of the adaptive landscape in a letter to Fisher in 1931 [46, 104]:

> Think of the field of visible joint frequencies of all genes as spread out in a multidimensional space. Add another dimension measuring degree of fitness. The field would be very humpy in relation to the latter because of epistatic relations, groups of mutations, which were deleterious individually producing a harmonious result in combination. (Wright to Fisher, February 3, 1931)

Wright was convinced that real adaptive landscapes must be complex due to pervasive epistasis. He imagined rugged landscapes with multiple adaptive peaks separated by low-fitness valleys: "it may be taken as certain that there will be an enormous number of widely

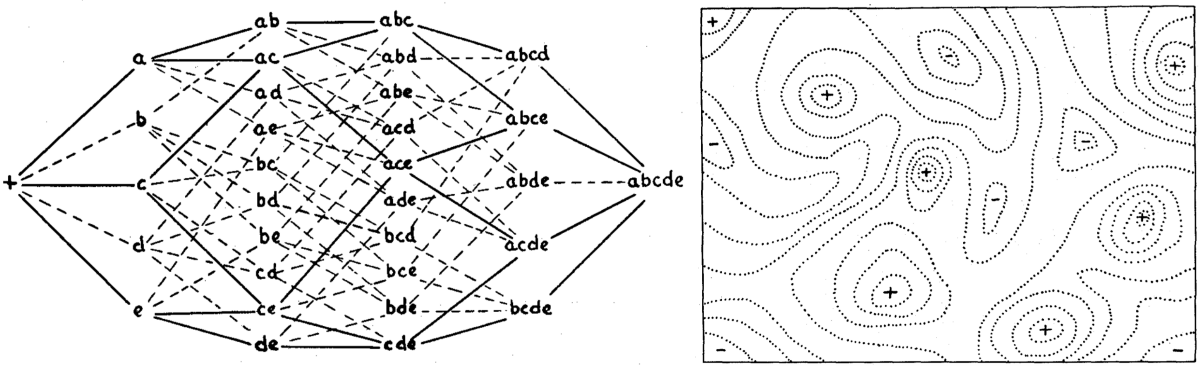


Figure 1.2: The adaptive landscape. An adaptive landscape is usually visualized as a "hilly" three-dimensional landscape where genotypes are arranged in the x–y plane, and fitness defines the "height" of the landscape on the z axis (e.g., Fig. 2.1). However, Sewall Wright did not use this representation in his seminal 1932 paper [40]. Instead, acknowledging full well the high-dimensional nature of genotype space, he used a genotype network representation. For example, if genotypes are defined by five loci with two alleles each (wild type and mutant), a network connects all genotypes from the wild type genotype (+) to the quintuple-mutant (abdce) connecting all five single-mutants, ten double-mutants, ten triple-mutants, and five quadruple-mutants (*left*). In his alternative representation of an adaptive landscape, Wright "compressed" this mutidimensional genotype space into a two-dimensional "field of gene combinations," where fitness is represented using contour lines (*right*). Due to pervasive epistasis, Wright envisioned a highly rugged landscapes with multiple high-fitness adaptive peaks (+) separated by low-fitness adaptive valleys (-). Images from [40], reproduced with permission of the Genetics Society of America.

separated harmonious combinations" [40]. However, Fisher doubted that this was the case because of his additive view of genetics [97]. Fisher thought that as the dimensionality of genotype space ("the field of gene combinations") increases, the number of adaptive peaks ("harmonious combinations") in the landscape should decrease (Fisher to Wright, May 31, 1931) [46, 104]. In other words, he thought that there is an inverse relationship between the number of adaptive peaks and the number of genotypic dimensions [106]. Therefore, according to Fisher, because real adaptive landscapes are high-dimensional, they should be single-peaked or have very few peaks.

1.1.3.2 Phenotypic landscapes

In 1944, the palaeontologist George G. Simpson—one of the fathers of the "modern synthesis" [107]—used adaptive landscapes to study macroevolutionary change [108]. Simpson's landscapes have been called phenotypic landscapes because their non-fitness dimensions represent phenotypic traits instead of genotypes [104]. In other words, phenotypic land-

scapes depict phenotype-fitness maps [109, 110]. Later, in 1966, the palaeontologist David M. Raup proposed the concept of the *theoretical morphospace*, as an extension of the adaptive landscape into the science of morphology [109, 111]. However, there are important differences between Wright's adaptive landscape and a theoretical morphospace. The different locations in an adaptive landscape are genotypes, but the locations in a morphospace are forms or morphologies. The non-fitness dimension in a morphospace are parameters in a geometric model of form, and instead of fitness, the "vertical" dimension is the frequency of a given form or morphology in nature. The palaeontologist George McGhee in his book "The Geometry of Evolution" gives several fascinating examples of this type of landscape [109], but I will focus here on work done by Raup himself, and especially by his graduate student John Chamberlain [109, 112–114], because to my knowledge, they studied the first empirical adaptive landscape.

In 1967, Raup constructed a morphospace for shell form in ammonoids [109, 112], which are an extinct group of swimming cephalopds related to the nautilus that still swims in today's oceans. Ammonoids thrived in the oceans of the Palaeozoic and Mesozoic with hundreds of species. Raup developed a mathematical model based on two parameters (W and D; the specific meaning of W and D is not relevant for my exposition) able to generate all possible "ammonoid-like" shell morphologies, including shells that may never have evolved. Raup plotted the frequency distribution of 405 ammonoid fossil species in W/D space (Fig. 1.3). He found a single peak, indicating a cluster of highly abundant forms. To explain why only a single peak was found, and in a specific location of the morphospace, a functional analysis was needed. It was Chamberlain, Raup's student, who performed such an analysis by conducting experiments to measure the drag coefficients of different ammonoid shells [109, 113, 114]. The drag coefficient of a shell form is inversely related to its swimming efficiency. Chamberlain's empirical data showed that there are two adaptive peaks in W/D morphospace, which corresponds to forms with the maximum swimming efficiencies (Fig. 1.3). Only one of these two peaks overlapped with the single peak in Raup's frequency morphospace. A developmental or an evolutionary constraint

could explain why one of the adaptive peaks is unoccupied [109]. Alternatively, there could be a trade-off with another functional property that renders the shells in the empty peak maladaptive [109]. However, a 2004 study using a larger dataset including 597 new ammonoid species showed that many species had indeed evolved shell forms in the region of the additional adaptive peak [109, 115]. This spectacular example shows the predictive potential of the adaptive landscape concept in evolutionary biology [54, 56, 110].

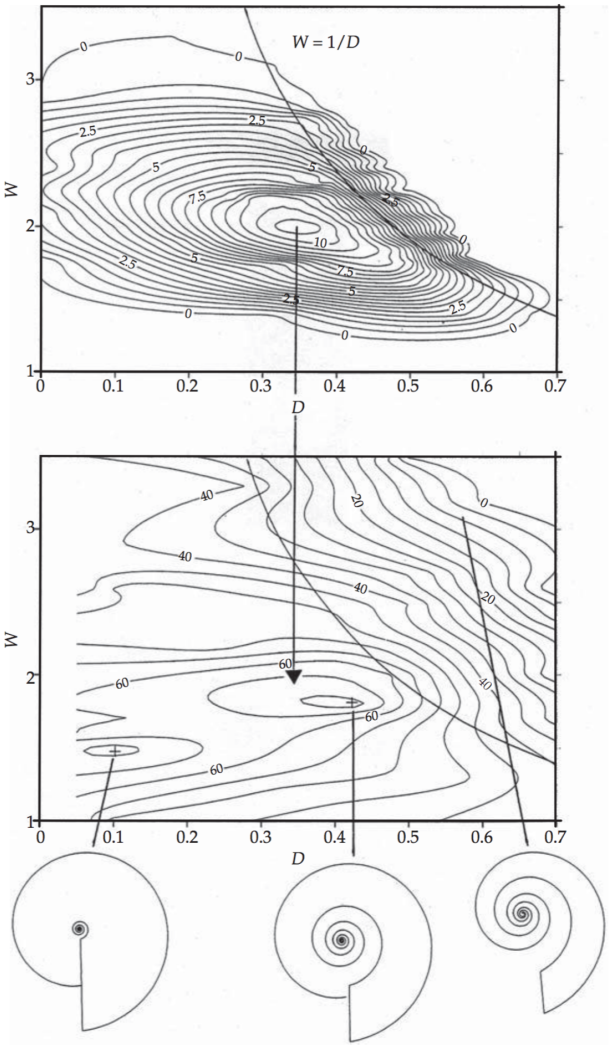


Figure 1.3: Phenotypic landscapes. The upper graph depicts, using contour lines, the frequency distribution of 405 actual ammonoid forms in Raup's theoretical morhospace [112]. There is a single frequency peak around $0.3 < D < 0.4$ and $W \sim 2$. The lower graph depicts the distribution of swimming-efficiency coefficients obtained by John Chamberlain [114]. There are two adaptive peaks in terms of swimming efficiency (+). One of them corresponds to the frequency peak in the top graph (arrow). Figure from [109], reproduced with permission from Cambridge University Press.

1.1.3.3 Molecular landscapes

In parallel to Raup's and Chamberlain's work, the concept of the adaptive landscape underwent a major development. After the birth and early development of molecular biology in the 1950s and 1960s, the relevant "genetic units" of genotype space were discovered [104, 116, 117]. With an improved understanding of the molecular basis of adaptation, the field of molecular evolution reframed evolutionary change in terms of molecular sequences (DNA, RNA, and protein) instead of alleles [104, 118–120]. It was in this context that John Maynard Smith imagined protein evolution as occurring in a protein space, that is, a space of all possible protein sequences [41, 121]. He envisioned a network of proteins where nodes are amino acid sequences, and where two nodes are connected by an edge if their sequences differ just by a single amino acid change. This network of genotypes is very similar to Wright's and Haldane's hypercube. To illustrate how proteins evolve in this space, Maynard Smith used a word game that requires to convert a word into another word of the same length by changing one letter at a time, and the requirement that all intermediate words must be meaningful in the English language (for example, changing "WORD" into "GENE" via "WORE," "GORE," and "GONE"). Similarly, functional proteins must evolve from other functional proteins through mutational pathways in the network of protein space, where every intermediate protein must be functional since natural selection would not favor a mutation into a non-functional protein. An essential requirement for this evolutionary process to work is that a certain proportion of single-mutant neighbors of a functional protein must also be functional [12, 41, 122]. In other words, functional proteins should display mutational robustness, that is, the ability to remain functional despite single point mutations [51].

In 1984, John Gillespie in the context of the neutralist-selectionist controversy proposed the metaphor of the mutational landscape [43, 104, 123, 124]. The mutational landscape is also a network in which nodes represent nucleotide sequences. Tow nodes are connected if their sequences are a nucleotide-substitution away form each other. In this landscape, mutational pathways are only accessible if every single mutation in the

pathway provides a selective advantage. Gillespie used the mutational landscape to model the rate of molecular evolution from a selectionist perspective.

1.1.3.4 Theoretical adaptive landscapes

Since Wright's introduction of the adaptive landscape in evolutionary biology, many theoreticians have studied how the topology of an adaptive landscapes influences evolutionary dynamics. Many models of adaptive landscapes have been developed over the years. An important type of such models are additive models, in which no epistatic interactions exist between different loci. The landscapes generated by such models are always single-peaked, a bit like Mount Fuji in Japan. Indeed, the most famous additive model is the so-called Mt. Fuji model [125]. In contrast, the House of Cards (HoC) model [126, 127], where the fitness values of any two genotypes are completely independent of one another, produces highly rugged landscapes, which are the opposite of the smooth landscapes produced by additive models. The Rough Mt. Fuji (RMF) model produces landscapes that still retain a global fitness maximum, that is, a global peak, but may contain lower peaks and plateaus [128, 129].

The theoretician Stuart Kauffman considerably advanced the study of how the topography of an adaptive landscape affects evolutionary dynamics when he developed the NK model, which is a model that allows the study of adaptation in landscapes with "tunable" ruggedness [96, 130–132]. In the NK fitness landscape model, N refers to the number of loci, and K to the number of epistatic interactions of each locus with other loci. In this model, it is possible to generate a spectrum of landscapes with different levels of ruggedness that are bookended by two extremes. One extreme is an additive Mt. Fuji-like landscape ($K = 0$). The other extreme is a highly rugged landscape with high levels of epistasis ($K = N - 1$). At $K = 0$, there are no epistatic interactions. There is a single high-fitness peak, with gentle slopes falling away from it. At $K = N - 1$, which is the maximum possible value of K, every locus interacts epistatically with all other loci, which results in a highly rugged landscape with multiple low-fitness peaks. Using NK landscapes, theorists have studied how landscapes ruggedness affects the number of local

peaks, the fraction of local peaks that can be accessed from a given genotype, and the average number of mutations needed to reach the global peak [131]. NK models have also been used to study RNA folding [133], how antibodies improve their affinity for an antigen during immune response [131], and regulatory circuits [96, 131].

Theoretical interest in adaptive landscapes goes beyond the frontiers of biology, since the problem of finding the global peak in an adaptive landscape is similar to the problem of finding global optima in multidimensional optimization problems. For instance, NK landscapes have been employed to study spin glasses in physics [96, 130]. Another example from computar sciences are evolutionary algorithms, which are population-based optimization algorithms that use mutation, recombination, and selection to find optimal solutions to different computational problems [134].

1.1.3.5 Empirical adaptive landscapes

At the advent of the twenty-first century our ignorance about the topography of real adaptive landscapes was still great. However, the development of high-throughput experimental technologies over the last decade is rapidly changing our understanding of landscape ruggedness in real living systems. It is now possible to construct empirical adaptive landscapes by measuring fitness or quantitative phenotypes for a large number of genotypes. Now, for the first time, large empirical landscapes can be comprehensively studied and analyzed. These new empirical landscapes are leading to novel insights into the structure of realistic landscapes and how it affects evolution [55–58, 135]. However, one of the main caveats of these studies is the high-dimensionality of genotype space, which Dobzhansky popularized with a now famous quote paraphrasing Wright [136]:

> Suppose there are only 1000 kinds of genes in the world, each gene existing in 10 different variants or alleles. Both figures are patent underestimates. Even so, the number of gametes with different combinations of genes potentially possible with these alleles would be 10^{1000}. This is fantastic, since the number of subatomic particles in the universe is estimated as a mere 10^{78}.

Stuart Kauffman has described such large numbers as hyper-astronomical because they

are many times larger than the gargantuan numbers used in astronomy [132]. For this reason, many empirical landscapes are constructed for single macromolecules rather than entire organisms. However, hyper-astronomical numbers also appear in macromolecules. The number of amino acid sequences for a protein with a hundred amino acids is 20^{100}, which is larger than 10^{130}. Hydrogen is the most common element in the universe, but the estimated number of hydrogen atoms in the universe is just 10^{90}, many order of magnitudes below the number of possible amino acid sequences for a protein of modest length [137]. Many local empirical adaptive landscapes of proteins [34, 52, 86, 88, 138–144] and RNAs [27, 31, 32, 145–147] have been studied. These landscapes are frequently incomplete, because they only assign fitness or phenotype to a minute fraction of genotypes that lie near a wild type genotype. In contrast, in chaper 2, I study more than a thousand landscapes which are complete as they describe transcription factor binding affinity to every single short DNA sequence of length eight.

A recent wave of studies on empirical adaptive landscapes has been made possible by technological developments, but it has been motivated by theoretical advances; especially by the discovery by Daniel Weinreich and co-workers that sign epistasis constrains evolutionary trajectories in adaptive landscapes [56, 148, 149]. Sign epistasis is a strong form of epistasis that changes the sign of the fitness effect of a mutation, from positive to negative or vice versa [148]. This type of epistasis reduces the number of mutational pathways in which fitness increases monotonically. These are also the pathways that are evolutionarily accessible under strong selection [148]. Sign epistasis is also necessary for the existence of multiple adaptive peaks [149], that is, the presence of several high-fitness genotypes that are surrounded by low-fitness genotypes. The fascinating prospect of predicting evolution emerged from the discovery of these evolutionary consequences of sign epistasis [56]. However, the quantitative study of this possibility requires, for a given pair of genotypes, the experimental construction of the 2^L genotypes that result from all possible combinations of the L mutations for which the two genotypes differ [56]. In a seminal study, Weinreich and colleagues constructed and analyzed an adaptive landscape of the β-lactamase TEM,

which is an enzyme that confers resistance against β-lactam antibiotics. They created all 32 (2^5) allele combinations of 5 mutations that jointly increase resistance to the antibiotic cefotaxime by 100,000-fold. Weinreich and co-workers measured the resistance provided by each of the 32 alleles as the minimal cefotaxime concentration inhibiting bacterial growth. The diagram of a genotype network with 32 genotypes in Fig. 1.2 depicts the kind of empirical landscape they constructed. In this landscape, they found that due to the prevalence of sign epistasis, only 18 of the 120 (= 5!) shortest mutational pathways from the wild type allele ("+" in Fig. 1.2) to the 5-mutant high-resistance allele ("abcde" in Fig. 1.2) were accessible under strong selection. In other words, cefotaxime resistance increases monotonically only in 15% of all 5-step pathways. And using predictions of fixation probabilities from classical population genetics, they showed that two of these pathways are more often taken by evolving populations than the other 16. Therefore, the evolution of cefotaxime resistance is highly predictable since it evolves most frequently along just a few mutational pathways. However, it is worth noticing that this landscape represents only a tiny fraction of all the $20^5 = 3,200,000$ possible combinations of amino acids at the five studied position of the β-lactamase TEM. The constraints imposed by sign epistasis on the evolution of cefotaxime resistance may differ in this larger landscape.

1.1.4 Genotype networks

Genotype networks [12, 69], or neutral networks [60], are graphs that contain as vertices all the genotypes that share the same phenotype, where vertices are connected by edges if their genotypes differ by a single mutation [41]. Such networks provide a visualization of a genotype-phenotype map that is complementary to low-dimensional representations, such as the classical three-dimensional rendering of an adaptive landscape. The reason is that a three-dimensional representation reduces the space of genotypes, which is multidimensional, to a two-dimensional plane. Consider, for example, that the genotype space for a protein of modest length 100, composed of 20^{100} genotypes, has already 100 dimensions. A low-dimensional representation is usually inadequate for providing more than a superficial understanding of genotype-phenotype maps. Genotype networks, on the contrary,

do not sacrifice the high-dimensionality of genotype spaces and are therefore more suited for a quantitative understanding of the structure of a genotype space.

Genotype space is organized as a hypercube graph where vertices represent genotypes, and two vertices are connected by an edge if they differ by a single mutation. As we have already seen (section 1.1.3.1), this was an insight that both Wright and Haldane obtained independently [40, 105]. However, the fact that this large graph can be pervaded by genotype networks—sub-graphs of genotype space—was only foreshadowed forty years later by Maynard Smith in his 1970 paper on protein spaces, that I already discussed in the context of the adaptive landscape (section 1.1.3.3) [41]. Maynard Smith speculated that "if evolution by natural selection is to occur, functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates" [41]. In a genotype network, all genotypes have the same phenotype. Maynard Smith considered a categorical binary phenotype: A protein genotype is associated with either a functional phenotype or a nonfunctional phenotype. In the genotype network envisioned by Maynard Smith all protein genotypes are functional.

Twenty years after this seminal paper by Maynard Smith, computational studies showed the existence of genotype networks in simple models of protein genotype-phenotype maps [59]. Specifically, David J. Lipman and W. John Wilbur studied a two-dimensional HP lattice model for protein folding with only two types of monomers—hydrophobic (H) and polar (P) amino acids [59]. It is the tendency of hydrophobic amino acids to avoid water molecules that drives protein folding. These binary sequences (protein genotype) fold into a two-dimensional lattice (protein phenotype), where each amino acid occupies a different position on a discrete grid. These authors found that usually a large number of genotypes fold into the same structure, and that these genotypes extend over large regions of sequence space, thus forming a genotype network.

One of the most detailed characterizations of genotype networks to date was performed with computational models for the *in silico* folding of RNA sequences into secondary structures through internal base-pairing. This work was carried out by Peter Schuster

and colleagues [60, 150, 151], and many others afterwards [69, 78, 84, 152–154]. Schuster *et al.* [60] first found that there are many different RNA sequences able to fold into the same secondary structure, and that these sequences form large genotype networks in sequence space. These authors coined these networks "neutral networks." In their definition of neutrality, a neutral mutation does not change the categorical phenotype associated with all genotypes in a neutral network. Therefore, these authors are using the term "neutral" with respect to a specific well-defined phenotype, knowing full well that such mutations may not be neutral with respect to fitness. However, in evolutionary biology, a mutation is neutral only if it does not affect fitness. Therefore, a neutral mutation with respect to fitness could be non-neutral with respect to a categorical phenotype, and vice versa. To avoid this confusion, Andreas Wagner introduced the alternative term "genotype network," which I use in this thesis [12].

Later, other authors found that genotype networks do not only pervade the genotype spaces of macromolecular systems such as proteins and RNA, but also appear in the genotype spaces of other biological systems of higher complexity, such as metabolism [62], and gene regulatory circuits [155]. Recently, genotype networks have been found using empirical data, thus validating a long-standing body of theoretical work on genotype-phenotype maps. Specifically, Payne and Wagner have discovered that the set of short DNA sequences bound by a transcription factor above some high-affinity threshold form a genotype network [156]. The phenotype in this case is the molecular ability of a site (genotype) to bind a specific protein factor. Specifically, they have discovered that for 99% of the studied factors, the majority of bound sites were part of a single connected genotype network. In other words it is almost always possible to transform one bound site into another via a series of mutations that preserve transcription factor binding. The mutational robustness of transcription factor binding sites explains the existence of these networks: The proportion of all possible single-mutants of a binding site that are also bound by the same transcription factor tends to be large. Genotype networks provide an entirely new way of analyzing protein-DNA interactions, one that has provided

new insights into transcriptional regulation systems, as we will see in chapters 2 and 3, where I study related genotype-phenotype maps. Additionally, this mapping of DNA sequences onto the proteins they bind constitutes the first exhaustive genotype-phenotype map entirely based on experimental data. Therefore, the study of Payne and Wagner [156] is not only the first time that genotype networks are used to analyze protein-DNA interactions, but also the first time they are applied to a genotype-phenotype map that is both empirical and comprehensive. In the next introductory section, I describe protein-DNA interactions in detail, and the high-throughput technology that has made these advances in studying empirical genotype-phenotypes possible.

1.2 Protein-DNA interactions: An empirical genotype-phenotype map

1.2.1 Transcription factors

The discovery of the *lac* operon and its regulation in *Escherichia coli* in the early 1960s by François Jacob and Jacques Monod showed for the first time the biological importance of protein-DNA interactions [116]. In a series of beautiful experiments these scientists discovered how the binding of a protein (Lac repressor) to DNA prevents the transcription of the lactose-metabolizing gene *lacZ* when no lactose is available in the environment. They were awarded the Nobel Prize in 1965 for this work. The Lac repressor is a transcription factor (TF), a sequence-specific DNA-biding protein that regulates gene expression by binding to DNA sequences known as TF binding sites [157, 158]. The binding of a TF to a gene's regulatory region may activate or repress the transcription of that gene by promoting or blocking the recruitment of RNA polymerase. The strength of this regulatory effect is partly determined by the TF's affinity for its site [159–162]. Genes coding for TFs typically represent 5-10% of the total number of genes in a given genome [163], and their products can regulate the expression of other TFs, forming transcriptional regulatory networks. Such networks control the development, behaviour, and physiology of many organisms, from bacteria to humans [164]. Accumulating evidence suggests that

many evolutionary adaptations can be explained by mutations in the regulatory regions of genes in these regulatory networks [165–167]. Such mutations, which change the timing or location of gene expression, have also been associated with human disease [168, 169]. Therefore, the characterization and study of the structure and function of regulatory networks constitutes an important and active area of research [170, 171], which critically depends upon our ability to measure and predict the affinity with which TFs bind their cognate sites.

TFs may have several functional domains—conserved protein segments that can function independently, each with a different function [158, 172, 173]. Although there are exceptions, they typically have just one DNA-binding domain, which can function autonomously. Other TF's domains are responsible for dimerization: Many TFs function as homodimers or heterodimers. Finally, some TF's domains mediate interactions with other proteins to form large molecular complexes that regulate the rate of transcription. For instance, many TF's have an activation domain that interacts with the basal transcriptional machinary to initiate transcription. TFs can be classified into families based on the structures and sequence similarity of their DNA-binding domains [158, 174]. TFs from the same family have similar structures, and thus bind DNA with the same overall geometry of interaction [158]. TFs from the same family also have a common ancestry, and have diverged through evolutionary processes such as gene duplication and species diversification.

1.2.2 Sequence-specific interactions between proteins and DNA

TFs have two major modes of specific interaction with DNA: Direct, and indirect [158]. Direct interactions are mediated by contacts between the lateral chains of the amino acids in a TF's DNA-binding domain and the edges of the base-pairs in the DNA. These contacts are both hydrogen bonds and van der Waals interactions. Most DNA-binding domain families interact with the major groove of DNA, although many domains also have

additional contacts in the minor groove. Different patterns of hydrogen-bond donors and acceptors and the methyl group on thymine allow TFs to discriminate between different DNA sequences. Indirect interactions are mediated by contacts between a TF and the DNA backbone, and depend on specific structural properties such as the width of the major groove, which can be both wider or narrower than in the standard B-form DNA depending on the sequence [175]. Because some sequences are prone to adopt specific structural deformations, indirect interactions are also sequence specific. Therefore, a specific DNA base-pair can establish a direct contact with the TF's DNA-binding domain, and at the same time contribute to the DNA shape in a way that favors indirect interactions. Finally, most TFs also have nonspecific contacts with backbone atoms. These contacts are electrostatic and provide a large component of the total binding energy. These nonspecific interactions are also important for the way in which TFs search for high-affinity binding sites by one-dimensional diffusion, that is, by "sliding" along the DNA until they find their sites.

The specificity of a DNA-binding protein is determined by its relative binding affinity to all possible binding sites. Some DNA-specific binding proteins (e.g., most restriction enzymes) bind specifically to a single sequence with high affinity. Transcription factors instead bind to many different sequences with varying binding affinities [158]. They usually bind to a preferred sequence that has the highest affinity, but other single-mutant neighbors of this sequence also have similar affinity. In some cases, sequences with multiple changes with respect to the highest-affinity sequence can also bind a TF with fairly high affinity [158].

The simplest way of describing the specificity of a DNA-binding protein is using the consensus sequence for its highest affinity sequences. However, the specificity of a TF is better described by using position weight matrices (PWMs). A PWM contains a score for each possible base (A, C, G, or T) at each of the positions in a binding site. A PWM allows the assignment of a score to every possible binding site. The score of a site is computed as the sum of the elements of the PWM that correspond to the base of the sequence at each

aligned position [158]. Usually, a threshold is used to classify if a site can be bound or not, such that sites with a score above the threshold are considered to be binding. Over the years, many computational methods have been developed to obtain PWMs from binding data. These statistical methods, including machine learning algorithms, take as an input a collection of sequences and their functional information obtained in high-throughput assays, such as those described in section 1.2.3 [176]. The most commonly used simple methods assume that each position in a binding site contributes additively to binding. However, there are also more complex methods that do not assume that each position contributes independently to binding affinity [177–183]. For example, using a PWM-like model that contains a score for each of the 16 possible dinucleotides at each position in a binding site allows for pairwise epistatic interactions between adjacent nucleotides [163].

1.2.3 High-throughput measurements of protein-DNA interactions

As a consequence of recent advances in microarrays and next-generation sequencing technologies, our ability to measure the affinity with which TFs bind DNA, both *in vitro* [184–187] and *in vivo* [188–191], has greatly improved. One of these new technologies, called a protein binding microarray (PBM) [184, 192, 193], measures the *in vitro* affinities of TFs to DNA sequences of up to k nucleotides in length, which are called k-mers. To date, this technology has been used to quantify the binding affinities of more than 1,000 TFs from 131 different eukaryotic species to all 32,896 possible 8-mers [194–196].

PBMs use double-stranded DNA (dsDNA) in a microarray format [184]. Single-stranded DNA (ssDNA) is placed on a spot of the array and then converted into dsDNA using a universal primer. Current arrays of over 44,000 spots contain all possible binding sites of 10 bases once in the array [184, 197]. This means that every 8-mer appears at least 16 times, and at least 32 times if they are nonpalindromic. The TF is added to the array, and the array is washed to remove nonspecific binding, and incubated with a fluorescent antibody (if the TF itself has not been fluorescently labelled). Comparing the

fluorescence intensity of every spot in the array allows the estimation of binding affinity to every dsDNA sequence of a given short length (up to length 10). For instance, PBMs serve to compute enrichment scores for all sequences 8-nucleotides long as a proxy for binding affinity [184, 197]. The score of an 8-mer is computed from the rank median fluorescence intensities of all the array sequences that contain the 8-mer relative to the average of the background. Cognate site identifier (CSI) methods are similar to PBMs. The main difference between these two technologies is that ssDNAs are designed in such a way as to fold back on themselves to produce dsDNA on the CSI array [185, 198, 199].

PBM data has provided us with two major observations that have enlarged our molecular and evolutionary understanding of TF-DNA interactions. First, TFs often bind multiple different sets of DNA sites with high affinity, while they can also bind hundreds of other sites with lower affinity. Second, the individual nucleotides of a TF binding site can contribute non-additively to binding affinity [200]. Namely, the contribution of one nucleotide may depend on other nucleotides in the site (epistasis). However, despite this recent progress, our knowledge about the relationship between epistasis and the evolution of TF-DNA interactions is still incomplete.

1.2.4 Epistasis in transcription factor binding sites

Epistasis has been frequently demonstrated in macromolecules such as RNA or proteins [52, 84–89]. However, whether the individual nucleotides that form TF binding sites contribute epistatically to binding affinity is still controversial [201, 202]. The debate mainly centers on the widespread use of methods for the prediction of TF binding sites based on PWMs that assume non-epistatic interactions between nucleotides [203, 204]. According to anecdotal experimental evidence, we now know that this additivity assumption is sometimes violated [186, 195, 200]. Yet the corrections introduced into these prediction methods merely consider pairwise non-additive interactions, and just between adjacent nucleotides [178]. However, epistatic interactions in TF binding sites may occur between multiple nucleotides that are not adjacent to one another [200]. In chapter 2, I comprehensively quantify the extent of epistasis in TF binding sites for more than 1,000 TFs.

1.3 Molecular chaperones: Modifiers of the genotype-phenotype map

A modifier of a genotype-phenotype map is a gene or genetic variant that modifies the phenotypic effect of other gene or variant via an epistatic interaction [205]. A global modifier influences many genes or variants at the same time. The concept of a modifier gene was introduced by Ronald Fisher to provide an evolutionary explanation to the phenomenon of dominance by which in diploid organisms a wildtype allele (+) dominates (hides) the phenotypic effect of a null allele (–) in the heterozygote genotype (+/–) [97, 206, 207]. Fisher proposed that dominance has evolved adaptively as robustness against mutations [51]. In his model, when a null allele first arises in a population, the phenotype of the heterozygote is intermediate to the phenotypes of both homozygotes (+/+ and –/–). Later on, dominance can evolve through selection on the modifier gene (which is different from the gene in which the null allele has emerged). Although Fisher's model has been criticized [51, 208–210], it has bequeathed the valuable concept of a gene that can influence the phenotypic variation segregating in a population—the variation upon which natural selection acts. In modern usage, a modifier gene is not necessarily a dominance modifier [205].

The most well-known modifier genes encode molecular chaperones [211]. Molecular chaperones are proteins that help other protein achieve their functional three-dimensional conformations. Chaperones are able to alter the mapping from protein genotypes into protein phenotypes [212]. Specifically, they increase the number of amino acid sequences that fold into the same structure, by buffering the negative effects of mutations that affect protein stability or folding. Therefore, they are proteins that can establish antagonistic epistatic interactions with many different deleterious genetic variants.

1.3.1 Molecular chaperones and protein misfolding

Molecular chaperones are present in all-three domains of life (eukarya, bacteria, and archaea), and thus are proteins that probably arose very early during the evolution of

primitive cells [213, 214]. Many chaperones are better known as heat-shock proteins (HSPs), because they are upregulated during stressful conditions (e.g. heat stress) in which aggregation-prone misfolded or unfolded proteins accumulate inside cells. Chaperones are typically named based on their molecular weight (HSP40, HSP60, HSP70, HSP90, etc.). The chaperones that participate in *de novo* protein folding and refolding belong to three major classes: HSP70s, HSP90s, and the chaperonins (HSP60s). They are large multigenic molecular machines that function through ATP-driven cycles of substrate binding and release, which are finely regulated by cofactors.

Chaperones increase the probability of a protein reaching its three-dimensional and functional native state via three main mechanisms. First, chaperones can bind to folding intermediates, preventing their aggregation and ensuing pernicious interactions with biological membranes and other proteins in the extremely crowded cellular environment. Second, binding or enclosing nascent proteins can narrow and smooth the folding landscapes that they have to explore, thus guiding them towards their native state [213]. Third, some chaperones can unfold misfolded proteins through an energetically costly process [215]. In other words, chaperones can act as unfoldases, that is, they can unfold misfolded proteins and refold them. Because protein synthesis has the highest energy cost inside a cell, protein misfolding is essentially a waste of energy. Chaperones can help save this energy [216].

Misfolding is deleterious not only because misfolded proteins are not able to perform their functions, but also because of the formation of protein aggregates (e.g., amyloid fibrils [217]), which are cytotoxic [218, 219]. Protein aggregates affect cell viability negatively, and are associated with important human diseases such as Alzheimer's disease and Parkinson's disease, in which these aggregates accumulate in brain cells [220, 221]. This fitness cost of protein misfolding may have left a trace in the genomes of every living organism. Specifically, for almost all sequenced genomes, there is a negative correlation between the proportion of preferred codons—the most abundant codon for each amino acid—in a gene (codon adaptation), and the gene's evolutionary rate, i.e., the

number of nucleotide substitutions per unit time [222–226]. There is also a positive correlation between protein expression level and codon adaptation, such that highly expressed genes evolve slowly, both in yeast [222] and *E. coli* [227]. It has been suggested that selection against mistranslation-induced protein misfolding could explain these correlations [224, 225]. Mistranslation is the process by which the ribosome incorrectly decodes mRNA, introducing phenotypic mutations into the synthesized protein [228–230]. In particular, selection against synonymous substitutions to maintain translation accuracy, and selection against amino acid substitutions to maintain protein robustness against mistranslation would explain why highly expressed genes evolve slowly and show high levels of codon adaptation [224]. Chaperones can ameliorate the impact of misfolded proteins by acting as unfoldases, and allowing polypeptides with destabilizing amino acids to fold into functional proteins.

The main physiological roles of chaperones include preventing protein aggregation, assisting protein folding and helping organisms to survive stressful conditions by restoring the native conformation of proteins destabilised by environmental perturbations. Through this last function chaperones provide a mechanism of environmental robustness—the resilience of a phenotype (e.g., cell fitness and protein stability or function) to environmental perturbations [51]. Chaperones also have the simultaneous capacity to buffer against deleterious mutations that affect the folding of proteins, that is, they also provide mutational or genetic robustness—the resilience of a phenotype to DNA mutations [51]. This last capacity has led to the claim that chaperones could act as evolutionary capacitors by promoting the accumulation of cryptic genetic variation that could be released during situations of environmental stress, thus facilitating rapid adaptive evolution [231]. This hypothesis was advanced based on studies with the chaperone Hsp90 (section 1.3.2). The bacterial chaperones most likely to be involved in this phenotypic buffering of genetic perturbations are DnaK and the chaperonin GroEL [232]. In chapter 4, I study DnaK as a source of mutational robustness, and in chapter 5, I study GroEL.

The protein folding pathways in the bacterial cytosol have been well-studied [213]. In

bacteria, the Trigger factor is the first chaperone to interact with nascent polypeptides. Most small proteins (~70% of total) rapidly fold after synthesis without further chaperone assistance. Longer proteins interact subsequently with the HSP70 system (DnaK–DnaJ) and reach their native conformation after several cycles of ATP-dependent binding and release (~20% of total). Finally, about ~10% of proteins require the chaperonin system (GroEL–GroES) to fold.

1.3.2 The molecular chaperone Hsp90

Hsp90 is a highly conserved molecular chaperone [233, 234]. The protein targets (clients) of this chaperone tend to be kinases, transcription factors, and ubiquitin ligases involved in cancer and signaling pathways. Selectivity is determined in part through interactions with dozens of co-chaperones. Previous work has demonstrated that inhibition of Hsp90—to the degree that it occurs in the context of some environmental stressors—can reveal the phenotypic effects of "cryptic" standing genetic variation in many different species [231, 234–238]. That is, changes in Hsp90 activity can transform the phenotypic impact of standing genetic variation. This suggests that Hsp90, by helping mutant proteins to fold, masks the fitness effects of mutations that would otherwise exert a phenotype. Even if Hsp90-buffered mutations are only rarely acquired, they could be enriched in a population if stabilizing selection does not remove them, because their deleterious phenotypic consequences are masked by the chaperone. In contrast, selection would more efficiently remove mutations which have immediate negative effects on fitness [239]. A recent study has found evidence for this hypothesis among mutations in yeast that affect cell size and shape [239].

1.3.3 The molecular chaperone DnaK

The bacterial chaperone DnaK belongs to the highly conserved HSP70 family [214]. It is a central player in the *E. coli* chaperone network responsible for protein folding and maintaining proteostasis (protein homeostasis). DnaK is highly abundant in the bacterial cytosol, where it interacts regularly with at least ∼700 mostly cytosolic proteins [240].

DnaK is expressed constitutively, and it is essential during heat stress (42 °C) and other types of cellular stress [240–243]. The ATP-driven reaction cycle of DnaK is regulated by DnaJ (HSP40 family) and the nucleotide-exchange factor GrpE. DnaJ determines the binding specificity of DnaK towards its protein targets (i.e., clients) [244, 245]. The chaperone system formed by these three proteins can both fold nascent polypetides (co- or post-translationally) and refold misfolded proteins [214]. It does so by binding to exposed hydrophobic patches—∼7 residues long and preferentially framed by positively charged amino acids—in unfolded or partially folded polypetides, thus preventing detrimental interactions with other polypeptides in the crowded cytosol [214, 246]. By successively binding and releasing a protein client in a cyclic process that consumes ATP, the chaperone system DnaK–DnaJ–GrpE allows the protein substrate to gradually explore its complex folding energy landscape [213, 214]. For some proteins, several of these bind-release cycles are enough to achieve the native conformation. However, other proteins that fail to fold after DnaK cycling can go into the folding chamber of the chaperonin GroEL [214].

1.3.4 The chaperonin GroEL

Chaperonins are large double-ring molecular complexes that globally enclose client proteins in a cylindrical folding chamber [213, 214]. Therefore, the size of client proteins is constrained by the size of such chamber, and typically oscillates between 20 KDa and 50 KDa. The GroEL–GroES system in *E. coli* has been extensively studied and is arguably the best-characterised chaperonin system of any organism. Both GroEL and GroES are expressed from the bacterial operon *groE*. GroEL belongs to the group I chaperonins (HSP60 proteins in eukaryotes), and works with the cochaperonin GroES (HSP10 proteins in eukaryotes), which acts as a lid for the folding chamber [214]. Specifically, GroEL is a homo-oligomeric complex integrated by two rings stacked back to back. Each ring is formed by seven monomers. Each GroEL monomer is composed of three domains separated by two hinge regions: the apical, intermediate and equatorial domains. The apical domain contains hydrophobic binding sites for unfolded or misfolded protein substrates and for some loops of GroES that allow its docking to GroEL. In the equatorial domain

near the hinge region there is a slot that binds ATP, whose hydrolysis moves GroEL forward through its reaction cycle, and provides a time window of ∼10 s for folding to take place [247]. That domain is also responsible for the interactions among the two rings of GroEL. GroES is a single heptameric ring that binds to one or both ends of the cylinder that is GroEL. It seems that GroEL employs the three mechanisms used by chaperones that I have mentioned above. It provides a protected folding environment that passively prevents aggregation and guides polypeptides in their folding through negative hydrophilic residues that cover the inside of its cylindrical chamber [248]. Additionally, recent evidence suggests that GroEL can unfold misfolded proteins through an ATP-directed stretching action [215, 247].

GroEL interacts with about 10% of cytosolic proteins in the *E. coli* proteome (∼250 different proteins). A subset of 50-85 of these proteins are considered obligate clients of the GroEL–GroES system, because they are absolutely dependent on that system for a proper folding [249, 250]. To explore the ability of GroEL to act as a general chaperone interacting with substrates that do not share any similarities in terms of sequence, structure or function, Wang *et al.* [251] used DNA shuffling to improve the capacity of GroEL to fold a specific substrate, the green fluorescent protein. However, that improvement came with the loss of its ability to fold a variety of natural substrates. This result shows that GroEL can display a substantial plasticity in terms of susbtrate specificity, but also reveals a conflict between specificity and generality.

1.3.5 Molecular chaperones as mutational buffers

Populations evolving under high mutational loads are prone to experience severe declines in fitness due to destabilising mutations in their proteins, which lead to high levels of misfolding. Bacterial endosymbionts of insects provide an example of this phenomenon [252–255]. These bacteria are maternally inherited by the host, and evolve under strong genetic drift due a population bottleneck they experience during their transmission to the host progeny. Because they are asexual, genetic recombination is not effective. Thus, bacterial endosymbionts evolve under a process similar to Muller's ratchet by which their

genomes accumulate deleterious mutations in an irreversible manner [256–258]. In these bacterial symbionts, GroEL tends to be highly abundant. It has been suggested that this high level of chaperonin expression helps to maintain protein stability under high loads of destabilizing mutations. This phenomenon was first observed in *Buchnera aphidicola*, an intracellular symbiont of aphids, in which the expression of the operon *groE* is 7.5 times higher than the wild-type expression level in its close free-living relative *E. coli* [259]. We know that GroEL is also naturally overexpressed in many other bacterial endosymbionts, and it is the most abundant protein together with the chaperone DnaK in these species [254, 260–265]. It is tempting then to suggest that chaperones in endosymbiotic bacteria are helping them cope with their high mutational loads via their buffering of destabilizng mutations.

This hypothesis has been put to the test experimentally. Mario Fares and colleagues evolved *E. coli* for ∼3,000 generations via single-cell bottlenecks [258]. This type of evolution experiment is known as a mutation accumulation experiment, because the bottlenecks reduce the efficiency of selection considerably, such that nonlethal mutations are free to accumulate under the influence of random genetic drift [44]. At the end of the experiment, as a consequence of their high mutational load, the evolved populations had roughly halved their fitness (growth rate). Overproducing GroEL in the evolved strains (∼86-fold higher abundance) restored fitness to ∼80% of the ancestral strains. However, this was only observed when supplementing the growth media with tryptone (a source of amino acids), which is probably required for the synthesis of large amounts of GroEL. This suggest that there is a trade-off between the benefits of chaperone buffering and the energetic cost of their production [266]. Similarly, Maisnier-Patin *et al.* [267] evolved hypermutable *Salmonella typhimurium* populations for ∼1,000 generations in a mutation accumulation experiment. Due to their mutational loads, the evolved populations showed lower fitness than the ancestral population. These evolved populations also showed levels of expression of DnaK and GroEL that were 2-3 times higher compared to the ancestral population. Additionally, a modest artificial 1.5-fold increase of GroEL improved fitness

substantially. These results complement the results by Fares *et al.* [258], providing further evidence that chaperones are a source of antagonistic epistasis that can mitigate the deletirious effect of accumulated mutations.

1.3.6 Molecular chaperones and protein evolution

The observation that molecular chaperones can act as mutational buffers, mitigating the negative fitness effects of high mutational loads prompted the idea that chaperones modulate the relationship between a protein's sequence and its final tertiary structure, that is, between genotype and phenotype [212]. As mentioned above, chaperones can stabilize proteins that have been destabilized by mutations, and in doing so, increase the number of primary sequences that can reach the same tertiary structure. Thus, chaperones can potentially expand the size of protein genotype networks in protein sequence space, and consequently increase protein evolvability [12, 268]. Genotype networks allow the accumulation of seemingly neutral, or hidden, "cryptic" genetic variation [269]. Both theoretical and experimental studies have shown that this type of variation, which is not expressed phenotypically, can accelerate adaptation to new environments [12, 51, 270–272]. However, the destabilizing effects of most mutations in proteins constrain the accumulation of cryptic variation [87, 273]. Indeed, over 80% of mutations affecting proteins are deleterious due to their destabilizing effects [274, 275]. Most mutations improving enzymatic activities are also destabilizing [251, 268, 276]. As we have seen above, destabilizing mutations are deleterious because they reduce functional protein levels due to misfolding, and can result in the formation of insoluble aggregates that reduce fitness [275, 277].

Nobuhiko Tokuriki and Dan Tawfik examined experimentally how GroEL can facilitate the accumulation of neutral genetic diversity and accelerate the rate of adaptive protein evolution [278]. They evolved several enzymes in multiple rounds of *in vitro* random mutagenesis followed by *in vivo* expression in *E. coli*. Every enzyme evolved under stabilizing selection either in the presence of GroEL overexpression, or in its absence. They observed that enzymes evolved under GroEL overexpression accumulated twice as many mutations, and these mutations had more than 3.5 times higher destabilizing effects

than in the absence of GroEL overexpression. Using the same experimental setting, they also performed adaptive evolution of a phosphotriesterase to increase its weak promiscuous esterase activity. They found that GroEL overexpression accelerated the rate of adaptation, and yielded more adaptive variants (≥2-fold), and variants with higher specificity and activity (≥10-fold) than under normal GroEL expression. These results show that the ability of chaperones to rescue stability-impaired mutants accelerates protein evolutionary rates.

After this experimental demonstration that chaperones buffering the phenotypic effect of deleterious mutations can facilitate adaptive evolution, further evidence was found in sequenced genomes. Because of chaperone buffering, the deleterious effect of destabilizing mutations in obligate chaperone clients should be lower than in sporadic clients, where they should be lower than in nonclients. Thus, the efficiency of purifying selection on purging deleterious mutations should decrease with increasing GroEL dependency, and protein evolutionary rates should increase with GroEL dependency. A comparison of *E. coli* proteins with their orthologs in other proteobacterial genomes revealed that this was the case [279, 280]. In chapter 4, I show similar results for the bacterial chaperone DnaK, which were published in 2016 [281]. In parallel, another study published in the same year also showed the acceleration of protein evolutionary rates due to DnaK buffering [282]. Taken together, these studies show how chaperone-mediated buffering accelerates the evolutionary rate of client proteins over long evolutionary time scales, and illustrate how an individual protein like a chaperone can have a disproportionate effect on genome evolution.

1.4 From metabolic genotypes to metabolic phenotypes

1.4.1 Metabolic networks

The metabolism of an organism is a large and complex system of chemical reactions organized in a highly reticulate reaction network [283]. Metabolic networks carry out two major biological functions essential for the maintenance of life [284]. The first is to

produce forms of energy useful to the organism from sources of energy present in the environment. The second is to produce biomass from nutrients in the environment that act as sources of chemical elements. Metabolic networks synthesize the small molecules required for cell growth from these nutrients. These include the proteinogenic amino acids, ribonucleotides, deoxyribonucleotides, lipids and enzyme cofactors. A free-living microorganism such as *E. coli* needs to synthetize more than 60 such biomass precursors in order to sustain growth [285, 286]. The reactions in a metabolic network are catalyzed by enzymes encoded by genes. Most of these enzymes are proteins. The number of reactions in a metabolic network varies from hundreds to thousands, depending on the complexity of the environment an organism experiences [285, 287]. However, in a given chemical environment not all reactions are essential for the synthesis of biomass precursors. These reactions can be deleted without abolishing the network's ability to sustain cell growth.

The study of the structure, function, and evolution of metabolic network is a highly active area of research with many decades of history [283, 284], and with rewarding applications in the fields of metabolic engineering and drug discovery and design [288–290]. Older studies focused on small networks with few reactions or linear pathways of reactions. However, in the mid-1990s, with the rise of systems biology and technological developments such as whole-genome sequencing, the focus shifted towards genomic-scale metabolic systems, which comprise most or all of the reactions in an organism's metabolism. Such genome-scale reconstructions of metabolism in many different organisms have revealed that metabolic networks tend to have a "bow-tie" architecture [291, 292]. Numerous pathways convert various nutrients into different metabolic precursor molecules that feed into a central core metabolism, which comprises several different biochemical pathways interlinked in a complex manner. From this metabolic core emerge many biosynthetic pathways that produce biomass precursors.

1.4.2 Flux balance analysis

One of the main goals of systems biology is to predict metabolic *phenotypes* from metabolic *genotypes*. This was precisely the research goal of J. Burns when he coined the term

genotype-phenotype map (section 1.1) [42]. A metabolic genotype is the set of enzyme-encoding genes in a genome, and a metabolic phenotype is the set of molecules a metabolic network can synthesize, as well as the rate at which it does so [62, 284]. At present, it is beyond our technological capabilities to experimentally map metabolic phenotypes from metabolic genotypes, despite great progress in the development of experimental techniques to study metabolism [293–295]. Thus, it is only possible to determine the phenotypes of genome-scale metabolic networks using computational approaches, such as, flux balance analysis (FBA) [296], which is a constraint-based method [297–300].

FBA requires information about all reactions in a metabolic network in the form of a stoichiometric matrix, $\mathbf{S}$, of size $m \times n$, where m represents the number of metabolites, and n the number of reactions. Each entry s_{ij} of $\mathbf{S}$ denotes the stoichiometric coefficient of metabolite i participating in reaction j. A negative coefficient indicates that a metabolite is being consumed in a reaction, while a positive coefficient indicates that a metabolite is being produced. Most of the coefficients in $\mathbf{S}$ are zero, indicating that a metabolite does not participate in a given reaction. In other words, $\mathbf{S}$ is a very sparse matrix because metabolic reactions involve just a small number of different metabolites. Metabolic fluxes are represented by a vector $\boldsymbol{v}$ of length n, whose entries, v_i denotes the rate of reaction i.

FBA assumes that a network is in a metabolic steady state, which imposes the mass conservation constraint by which all internal metabolites are synthesized and consumed at the same rate. Such a metabolic state may be found in microbial cells growing exponentially in a constant environment (e.g., chemostat). The constraint of mass conservation on the network's metabolites can be formalized mathematically as:

$$\mathbf{S}\boldsymbol{v} = \mathbf{0} \tag{1.1}$$

Because there are more unkown variables than equations—there are more reactions than metabolites, $m < n$—there are many different vectors $\boldsymbol{v}$ that satisfy this equation. Together, they form a large allowable solution space (the null space of $\mathbf{S}$).

In order to restrict this solution space to biologically meaningful solutions, FBA uses

linear programming to maximize certain phenotypic properties, such as ATP synthesis, or the rate at which biomass precursors are produced in balanced amounts. The objective function that FBA maximizes is $Z = \boldsymbol{c}^{\intercal}\boldsymbol{v}$, where $\boldsymbol{c}$ is a vector of weights in which the element c_i indicates how much reaction i contributes to the objective function. Additionally, vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ contain upper and lower bounds on the fluxes of each reaction in the network. That is, the flux of reaction i has an upper limit a_i and a lower limit b_i. Besides the constraint $\mathbf{S}\boldsymbol{v} = \mathbf{0}$, these limits also contrain the maximization of Z, and reduce the number of allowed flux distributions $\boldsymbol{v}$. The output of FBA is a flux vector $\boldsymbol{v}$ that maximizes Z given these constraints.

1.4.2.1 Monte Carlo sampling of flux space

The solution space of FBA typically contains many conceivable steady-state flux distributions that satisfy the imposed constraints. Equivalent optimal solutions can exist in this space [301]. For instance, if two parallel pathways can produce the same metabolite at the same maximal rate, it is impossible to distinguish without experiments which one is being used [302]. Monte Carlo sampling provides a way to obtain a set of feasible flux distributions (points in solution space) in an unbiased manner, and it can be applied to large genome-scale metabolic networks [302]. The most commonly used algorithm is an improvement to Markov chain Monte Carlo (MCMC): artificial centering hit-and-run (ACHR) [303–305]. Sampling starts with a set of "warm-up points," which are non-uniform pseudo-random points within the high-dimensional solution space. Each of these points is a valid solution. Then, each warm-up point moves in a random walk within the limits of the solution space. First, a direction is chosen randomly for each point. Second, a limit to how far a point can move in this random direction is calculated, also randomly. Lastly, a new random point is selected along the line of movement of the previous point. This process is iterated many times until the set of points approaches a uniform "well-mixed" sample of the solution space. The uniformity of the sample can be estimated using a metric known as mixed fraction [304, 306]. Specifically, a line is defined such that half of the warm-up points are on either side of the line. The mixed fraction

is just the fraction of points that cross the line at the end of the sampling procedure. At the beginning the mixed fraction is 1, but at the end it approaches 0.5 if the sample is uniform. In summary, Monte Carlo sampling provides for each metabolic reaction a distribution of flux values. Each of these distributions inform us about the range of flux values a reaction can take, as well as about the probability of a given flux value [305]. Some reactions are really constrained and have a narrow distribution of flux values, while others can take a wide range of flux values given the model constraints and the topology of a metabolic network [307].

1.4.3 Evolution of metabolic networks

Natural selection acting on a metabolic phenotype constrains the evolution of a network's enzymes [284, 308]. Conversely, the evolution of enzymes through gene duplication, gene deletion, regulatory mutations, and amino acid substitutions can affect the networks's phenotypes [12, 284]. The first of these two perspectives on the evolution of metabolic networks is the topic of chapter 6, where I study the evolution of enzyme-encoding genes through point mutations. Point mutations affecting the coding region of a metabolic gene can change the activity of the encoded enzyme, for example by changing its catalytic efficiency. Such mutations can modestly increase or reduce the activity of an enzyme, but they can have large effects on the fitness of a cell [309, 310]. They can even allow growth on a new carbon source [311].

Over evolutionary time, a microbial metabolism can evolve very rapidly through changes that involve the loss or gain of genetic material. A metabolic network can lose enzymes through gene deletion or loss-of-function mutations, or acquire new ones through horizontal gene transfer or gene duplication. Horizontal or lateral gene transfer occurs in both prokaryotes and eukaryotes, but it is so rampant in prokaryotes that it can modify their genomes on short time scales [12, 312]. For example, different *E. coli* strains may differ by more than 20% of their genomes, and may have gained around 100 genes via lateral gene transfer relative to other strains [12]. On average, they differ in 36% of the reactions in their metabolic networks [313]. Newly acquired enzymes may catalyze

new chemical reactions conferring novel metabolic phenotypes onto a metabolic network. Another mechanism for the evolution of new enzymes is gene duplication, although horizontal gene transfer is a greater contributor to the recent evolution of bacterial genomic repertoires [314].

1.5 Thesis outline

In my thesis, I study several genotype-phenotype maps for different complex living systems, using a combination of experimental and computational approaches (Table 1.1). In chapter 2, I study genotype-phenotype landscapes, where each genotype is a short DNA binding site, and the phenotype is a quantitative measure of binding affinity for specific TFs. In chapter 3, I study a similar genotype-phenotype map, where each genotype is also a short DNA binding site, but the phenotype instead of being quantitative is categorical, i.e. the site's capacity to bind specific TFs (or members of specific DNA binding domain families). These are biologically important phenotypes, because TF binding is integral to the transcriptional regulation of gene expression, which underlies fundamental developmental, behavioral, and physiological processes. In chapters 4, and 5, I study how bacterial chaperones alter the mapping from protein genotypes to protein phenotypes, i.e. into the structures that proteins form. Specifically, I study the evolutionary consequences of this modified protein genotype-phenotype map on genome evolution. Finally, in chapter 6, I analyze the mapping from metabolic genotypes—a genome's set of enzyme-coding genes—to metabolic phenotypes—the set of molecules a metabolism can synthesize, and the rate at which it does so. Specifically, I study how selection for a given metabolic phenotype can constrain enzyme evolution. Here is a brief summary of these studies.

In chapter 2 [74], I study a large number of complete and empirical adaptive landscapes. To accomplish this, I focus on transcriptional regulation, which drives the development, behavior and physiology of organisms as different as bacteria and humans. More specifically, I focus on the interaction between TFs and their DNA binding sites. It is well known that evolutionary change in TF binding sites is both an important means

Table 1.1: Genotype-phenotype maps studied in this dissertation.

Genotype	**Phenotype**	**Type**[a]	**Chapter**
Transcription factor binding site	Binding affinity	Q	2
Transcription factor binding site	Transcription factor	C	3
Transcription factor binding site	DNA-binding domain	C	3
Amino acid sequence	Protein structure	C	4,5
Metabolic genotype	Metabolic phenotype	C	6

[a] Type of phenotype: C: Categorical; Q: Quantitative. See section 1.1.1.

by which gene regulation has evolved and a common culprit in disease. Such change is also involved in countless evolutionary adaptations. This study benefits from high-throughput technologies developed in recent years, such as protein binding microarrays and digital footprinting that have facilitated the characterization of thousands of potential TF binding sites in many different organisms, both *in vitro* and *in vivo*. Specifically, I characterize and study 1,137 adaptive landscapes of transcriptional regulation from 129 different eukaryotic species. Each landscape is derived from protein binding microarray data, and describes the binding affinity of a TF to all possible binding sites. In these genotype-phenotype landscapes, adaptation is the exploration of sequence space that tries to optimize the capacity of a short DNA sequence to bind a particular TF in the absence of confounding factors, such as chromatin context. The data I study bring the metaphor of an adaptive landscape to life at unprecedented resolution. To my knowledge, this is the first time that multiple complete, empirical adaptive landscapes have been characterized, providing an exceptional opportunity to study their topography and navigability through single nucleotide changes. Affinity-modulating mutations in TF binding sites are important drivers of adaptation, and this study provides fundamental insights into how such mutations may fine-tune binding affinity. In doing so, it sheds further lights on the evolvability of transcriptional regulation.

In chapter 3, I build upon a large and long-standing body of work that has sought to elucidate the architecture of genotype-phenotype maps for categorical phenotypes. Until recently, this endeavor has been restricted to computational models of biological systems, due to a lack of experimental data. Here, I go beyond the state of the art by

studying a global, high-resolution depiction of an empirical genotype-phenotype map for a large number of categorical phenotypes from transcriptional regulatory systems. In this map, genotypes are DNA sequences and phenotypes are the TFs that bind these sequences. I study this genotype-phenoptype map using genotype networks, in which nodes represent genotypes with the same phenotype, and edges connect nodes if their genotypes differ by a single small mutation. These networks have been used before to study the relationship between robustness and evolvability in TF binding sites [156]. Here, I extend this earlier work by describing the structure and arrangement of these networks within the space of all possible binding sites for 527 TFs from three eukaryotic species encompassing three kingdoms of life (animal, plant, and fungi). In summary, I provide a global and detailed characterization of this genotype space for hundreds of TFs, thus describing the architecture of an empirical genotype-phenotype map at high resolution.

In chapter 4 [281], I study how an individual molecular chaperone such as the bacterial DnaK protein, through its role in protein folding, can have a disproportionate effect on the evolution of a proteome. I analyze evolutionary rates of proteins that are subject to DnaK-assisted folding on short, intermediate, and long evolutionary time scales through a combination of experimental and comparative approaches. Most of the evidence I find indicates that DnaK can buffer mutations in its client proteins, and that these proteins therefore evolve faster than in the absence of DnaK-mediated folding. This is the first demonstration that a member of the Hsp70 family of chaperones can buffer the effect of destabilizing mutations, with long-term consequences on protein evolution.

In chapter 5, I study how hypermutable *E. coli* populations evolved through single-cell bottlenecks for thousand of generations decrease their ability to metabolize most carbon substrates. I find that overproduction of the chaperone GroEL can ameliorate this metabolic erosion in some environments, likely because of its buffering of destabilizing mutations in metabolic enzymes.

In chapter 6, I study how the structure and function of a large bacterial metabolic network influences the evolution of its constituent enzymes. My analysis is part of a

research tradition aiming to understand the molecular evolution of living systems by relating the evolutionary rates of genes with their function and position in a biological network. An advantage of using metabolic networks instead of protein-protein interaction networks, which are more commonly used in this type of studies, is that the relationship between the functions of the enzymes and the network is especially well understood. To my knowledge, this is the first time that such a study is performed using the whole-genome metabolic reconstruction of *E. coli*, which is arguably the best-known metabolic network of any living organism. Specifically, I study how quantities such as enzyme connectivity and metabolic flux—the rate at which a reaction transforms substrates into products—affect evolutionary rate. To do so, I account for possible flux variation with Markov chain Monte Carlo sampling, a method that has not been used before in this type of evolutionary analysis. Additionally, I also study for the first time the influence of factors such as reaction superessentiality, which quantifies how easily a reaction can be bypassed in a metabolic network by other reactions or pathways, and the number of different chemical reactions that an enzyme catalyzes. In performing these analyses, I comprehensively characterize metabolic determinants of enzyme evolution, and show that an enzyme's role in the function of a metabolic network affects its evolution more than its place in the network's structure. In doing so, I illustrate how a systems-level perspective can help understand the factors that contribute to protein evolution.

2 A thousand empirical adaptive landscapes and their navigability

Published as:

Aguilar-Rodríguez J, Payne J L, and Wagner A (2017) *Nature Ecology & Evolution*, 1: 0045.

Abstract

The adaptive landscape is an iconic metaphor that pervades evolutionary biology. It was mostly applied in theoretical models until recent years, when empirical data began to allow partial landscape reconstructions. Here, we exhaustively analyse 1,137 complete landscapes from 129 eukaryotic species, each describing the binding affinity of a transcription factor to all possible short DNA sequences. We find that the navigability of these landscapes through single mutations is intermediate to that of additive and shuffled null models, suggesting that binding affinity—and thereby gene expression—is readily fine-tuned via mutations in transcription factor binding sites. The landscapes have few peaks that vary in their accessibility and in the number of sequences they contain. Binding sites in the mouse genome are enriched in sequences found in the peaks of especially navigable landscapes and the genetic diversity of binding sites in yeast increases with the number of sequences in a peak. Our findings suggest that landscape navigability may have contributed to the enormous success of transcriptional regulation as a source of evolutionary adaptations and innovations.

2.1 Introduction

An adaptive landscape is a mapping from a high-dimensional space of genotypes onto fitness or some other related quantitative phenotype, which defines the 'elevation' of each coordinate in genotype space [40]. Evolution can be viewed as a hill-climbing process in an adaptive landscape, where populations tend to move towards peaks as a consequence of natural selection. The ruggedness of an adaptive landscape has important evolutionary consequences, particularly for the evolution of sex, reproductive isolation and mutational robustness, and for the predictability of evolution [55]. An adaptive landscape that is smooth and single peaked does not pose any obstacle to evolutionary exploration. It is therefore highly navigable, in that it is possible to reach the global peak via positive selection through a series of small mutations that only move 'uphill'. In contrast, a rugged landscape can block the approach to the highest peak by entrapping populations on local suboptimal peaks [96].

We know very little about the navigability of empirical adaptive landscapes, largely due to the incompleteness of the landscapes that have been constructed to date. With few exceptions [25, 27], these landscapes were built by assaying the phenotypes of only a small number of mutations in all possible combinations within a single wild-type background [55]. These studies have helped form our intuition about the structure and navigability of empirical adaptive landscapes, but their conclusions are limited by the fact that they describe only a minute fraction of any complete landscape. An additional caveat of earlier studies is their focus on just one or a few landscapes, which limits the generality of their findings.

To study the navigability of a large number of complete, empirical adaptive landscapes, we consider data that describe the binding affinity of a transcription factor (TF)—a sequence-specific DNA-binding protein that helps regulate gene expression—to all possible DNA sequences (TF binding sites) of eight nucleotides in length. TFs are fundamental mediators of gene expression and are involved in numerous evolutionary innovations [315]. Their regulatory effect can be modulated via mutations in TF binding sites, which may

alter a TF's affinity for a site and thereby affect gene expression [160, 190, 316]. We describe the mapping of DNA sequence to binding affinity as an adaptive landscape, where we can study selection for TF binding. This is a common approach for exploring the evolution of TF binding sites [186, 317–320], other protein-DNA interactions [25, 199, 321] and protein-RNA interactions [28]. In this context, adaptive evolution is an exploration of sequence space that attempts to optimize the capacity of a sequence to bind a particular TF.

2.2 Results

2.2.1 Adaptive landscapes of TF binding affinity

We obtained protein- binding microarray (PBM) data for 1,137 TFs from the UniPROBE [322] and CIS-BP [196] databases. These TFs represent 129 eukaryotic species and 62 different DNA-binding domain structural classes. For each TF, we construct an adaptive landscape from the enrichment score (E-score)—a proxy for relative binding affinity [156, 184, 195]—of each of the 32,896 possible sites that bind the TF ('Methods'). These landscapes are complete because they describe the affinity with which a TF binds all possible sites, in the absence of confounding factors such as epigenetic marks, chromatin context, local sequence context or interactions with protein partners. We consider a sequence as 'bound' if its E-score exceeded 0.35 [156, 194, 195, 323] ('Methods'). We use this binding affinity threshold (τ) to differentiate sequences that are specifically bound by a TF via hydrogen-bond donors and acceptors from those that are unspecifically bound by a TF, for example, via its affinity for the DNA backbone. To facilitate the analysis of these landscapes, we represent each of them as a genotype network [41], in which vertices represent bound DNA sequences and edges connect sequences that differ by a single point mutation or a short insertion/deletion [156] (Fig. 2.1a; 'Methods'). These networks sometimes comprise multiple disconnected components (Supplementary Section 2.5.1 and Supplementary Figs 2.1– 2.3); when this occurs, we consider only the largest component, which we refer to as the dominant genotype network. Each dominant genotype network

forms the basis of an adaptive landscape, in which binding affinity defines the 'elevation' of each coordinate (TF binding site) in genotype space.

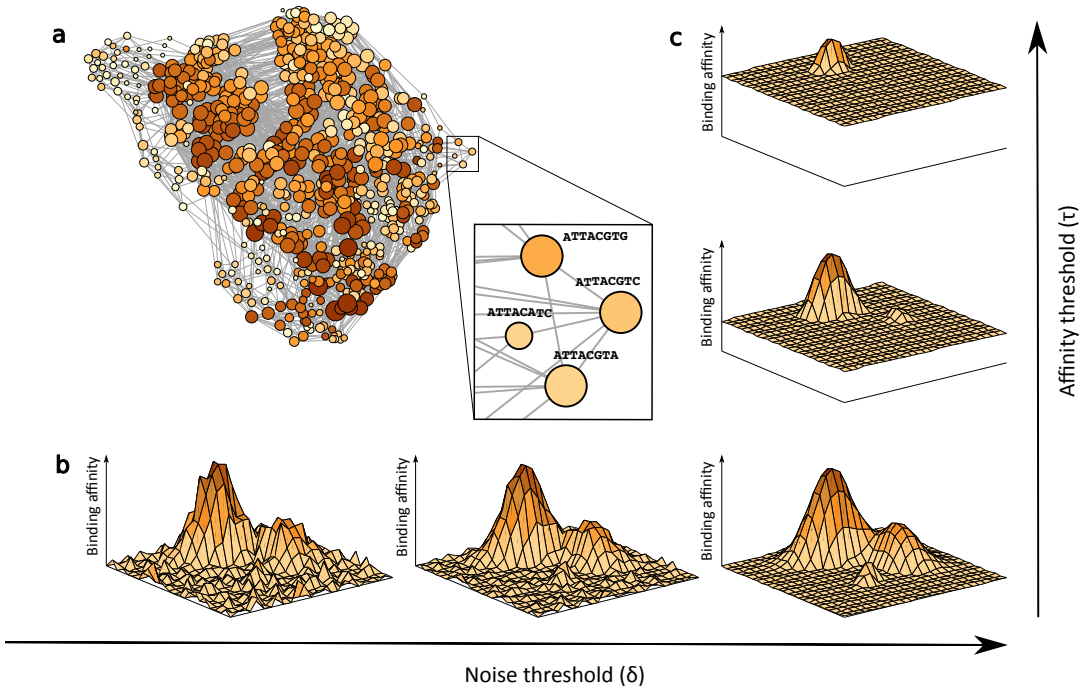


Figure 2.1: Adaptive landscapes of TF binding affinity. **a**, The largest connected component of the genotype network for the yeast TF Gcn4, visualized using a force-directed algorithm. Each vertex corresponds to a DNA sequence bound by Gcn4 (E-score > 0.35). The colour of a vertex indicates its binding affinity (darker = higher), that is, the 'elevation' of the landscape, whereas vertex size corresponds to the number of neighbouring sequences (bigger = more). The inset shows that two vertices are connected by an edge if their corresponding sequences are separated by a single small mutation ('Methods'). **b**, To determine whether two sequences differ in their binding affinity, that is, if they differ in their 'elevation', we used a noise threshold, δ, which accounts for experimental noise in the PBM data. Increasing δ increases the 'smoothness' of the landscape, as shown in the three hypothetical landscapes at the bottom of the figure, where the x–y plane represents genotype space, and the z axis represents binding affinity. **c**, To delineate bound from unbound sequences we used a binding affinity threshold, τ. As τ increases, the number of bound sequences decreases, thus reducing the size of the landscapes and pruning local peaks, as shown in the three hypothetical landscapes at the right of the figure. In Supplementary Sections 2.5.6.1 and 2.5.6.2, we assess the sensitivity of our main results to broadly varying δ and τ values.

We define landscape navigability as the ability to access a global peak via an evolutionary exploration involving random mutation and natural selection. Landscape navigability is highest when all mutational paths to the global peak exhibit a monotonic increase in binding affinity, which implies a landscape that is smooth and single peaked.

Landscape navigability is lowest when no mutational paths to the global peak exhibit a monotonic increase in binding affinity. This implies a rugged landscape with many peaks. Our measures of landscape navigability depend on two parameters, the noise threshold (δ) and τ, which we use for noise filtering and for delineating bound from unbound sequences, respectively (Fig. 2.1b,c; 'Methods'). We compare the navigability of the 1,137 empirical landscapes to landscapes generated via two different null models that provide lower and upper bounds on navigability ('Methods'). In the first null model (the additive model), we deterministically assign a binding affinity to each of the TF binding sites in the genotype network using the position weight matrix (PWM) of the TF, which assumes additive interactions between nucleotides and therefore produces smooth and highly navigable landscapes. In the second null model (the shuffled model), we randomly permute the affinities of the TF binding sites in the genotype network, which yields a rugged landscape that hinders navigability. Due to the stochastic nature of the shuffled model, we generate 1,000 shuffled landscapes for each TF. These null models capture two opposing extremes of landscape navigability while maintaining the structure of the underlying genotype network and therefore provide two points of reference for the empirical landscapes.

2.2.2 Landscape navigability: the number of peaks

The number of peaks in an adaptive landscape is an important indicator of its navigability. The more peaks a landscape has, the less navigable it becomes, if the peaks are of unequal height. We find that 42% of the empirically derived landscapes (478 of 1,137) have multiple peaks of unequal height (Supplementary Figs 4 and 5), with peak numbers ranging from 2 to 36 (Fig. 2.2a). In comparison, only 4.5% of the additive landscapes (51 of 1,137) have multiple peaks, whereas 99% of the shuffled landscapes (1,125,800 of 1,137,000) have multiple peaks (Fig. 2.2a). One might think that larger landscapes with more binding sites also contain more peaks and earlier theoretical work hints at that possibility [96]. However, the experimental data we analyse show no such scaling relationship (Supplementary Fig. 6a). Thus, by the criterion of peak number, smaller landscapes are not necessarily more

navigable than larger landscapes. Also, while one usually thinks of a peak as a single sequence, the data do not support this notion. Except for 36 landscapes, the global peaks—those containing the highest-affinity site—are plateaus containing between 2 and 121 sequences (Supplementary Section 2.5.2 and Supplementary Figs 2.5 and 2.7).

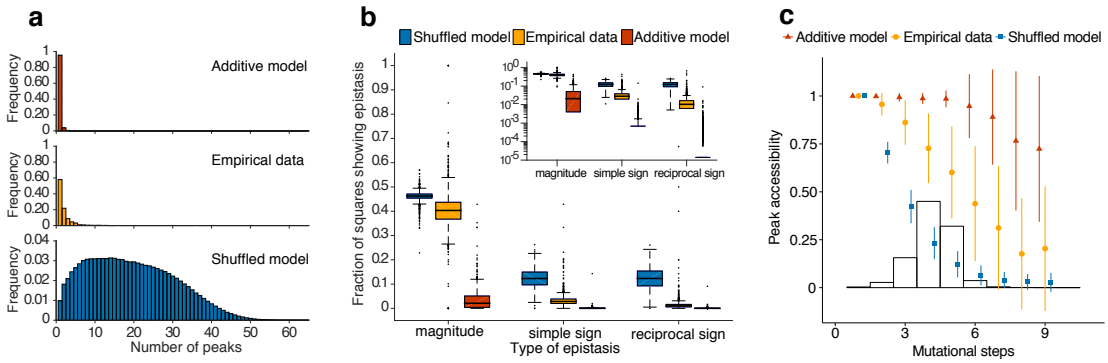
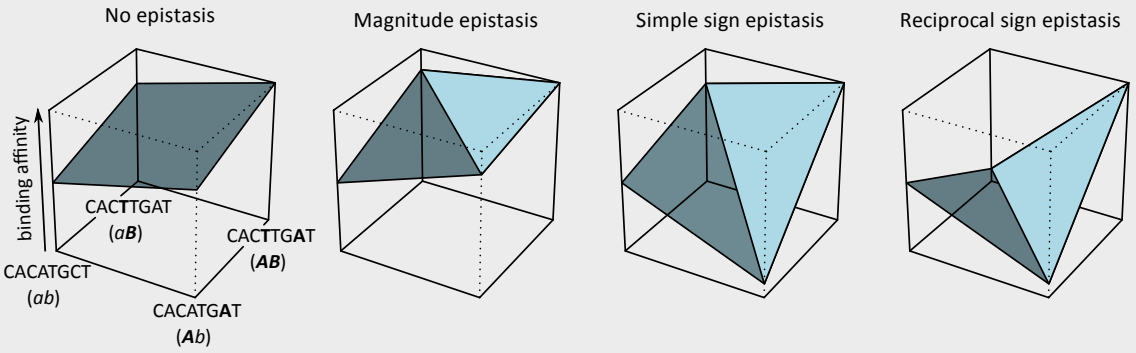


Figure 2.2: The navigability of adaptive landscapes of TF binding affinity. **a**, The distribution of the number of peaks for the 1,137 empirical adaptive landscapes (yellow) and the additive (red) and shuffled (blue) null models. **b**, Boxplots of the fraction of squares showing magnitude, simple sign and reciprocal sign epistasis for the 1,137 adaptive landscapes. The thick horizontal line in the middle of each box represents the median of the data, while the bottom and top of each box represent the 25th and 75th percentiles, respectively. For the shuffled model, the boxplot summarizes 1,137 data points, each of which is an average of 1,000 shuffled landscapes. All pairwise differences are significant (paired t-tests: P value $< 2.2 \times 10^{-16}$). The inset shows the same data, but with a logarithmically scaled y axis, which obscures the following numbers of data points with zero epistasis: magnitude epistasis (empirical, 9; additive, 184), simple sign epistasis (shuffled, 2; empirical, 53; additive, 681) and reciprocal sign epistasis (shuffled, 1; empirical, 98; additive, 853). The absence of the thick horizontal line indicates that the median of the distribution is below the lowest value of the logarithmically scaled y axis. **c**, For each of the 1,137 adaptive landscapes we show the mean (symbols) and standard deviation (error bars) of the fraction of accessible paths to the highest- affinity site in the landscape (that is, the peak accessibility). The histogram shows the distribution of mutational steps to the highest-affinity site for all sequences in all landscapes. For the shuffled model, each symbol represents the mean of 1,137 data points, each of which is an average of 1,000 landscapes.

Box 1: Epistasis

A square in a genotype network connects a 'wild type' sequence (ab) to a double mutant (AB) through two single mutants (Ab and aB). The left-most panel shows an example with binding sites of length eight where there is no epistasis: the binding affinity of the double mutant is simply the addition of the affinity contributions of the two single mutants. That is, the mutation a to A has the same effect on binding affinity in the two different genetic backgrounds (b and B). Magnitude epistasis occurs when the magnitude (but not the sign) of a mutation's effect on binding affinity depends on the genetic background. Simple sign epistasis occurs when one single mutant has a lower binding affinity than both the wild type and the double mutant, while the other single mutant has an affinity that is intermediate to the wild type and double mutant [148]. Reciprocal sign epistasis occurs when both mutations decrease affinity independently, but increase affinity in combination [149]. To account for the noise in our data, we only considered a single mutant's binding affinity to be lower than that of the wild type if the affinity difference exceeded a threshold value δ, which was derived from the empirical data and was specic to each TF ('Methods'). For all squares, we quantified epistasis along a single axis of the square by designating the highest-affinity sequence as the double mutant ('Methods'). Figure adapted from ref. [82], Macmillan Publishers Ltd.



2.2.3 Landscape navigability: epistasis

Epistasis [83]—non-additive effects of different mutations on a quantitative phenotype (or fitness)—can increase the ruggedness of an adaptive landscape. In the absence of epistasis, adaptive landscapes are smooth and single peaked, and thus do not hinder evolutionary

exploration. Epistasis can be partitioned into three different classes—magnitude, simple sign and reciprocal sign—with increasingly detrimental effects on landscape navigability (Box 1) [149]. To study epistasis, we first identify all squares in a genotype network—a binding site, two of its one-mutant neighbours and the double mutant that can be formed from the single mutants—and classify their affinity relationships according to the scheme in Box 1 ('Methods'). In the additive landscapes, on average, no more than 0.1% of squares show either of the two kinds of epistasis that impede landscape navigability most severely (simple or reciprocal, collectively referred to as sign epistasis) (Supplementary Section 2.5.3 and Supplementary Fig. 2.8). In the experimental data, sign epistasis is more frequent and affects 4.7% of squares, on average (Fig. 2.2b). However, its incidence is still five times lower than in the shuffled landscapes, where it affects 24.5% of the squares. In the experimental data only magnitude epistasis, which does not affect landscape navigability, approaches the levels observed in the shuffled landscape. In addition, sign epistasis preferentially occurs among nearby nucleotides in a binding site [324], whereas magnitude epistasis shows no such preference (Supplementary Section 2.5.4 and Supplementary Fig. 2.9).

2.2.4 Landscape navigability: accessible mutational paths

Another important indicator of landscape navigability is the fraction of accessible mutational paths to a given genotype from all other genotypes in the landscape (Supplementary Fig. 2.10). Here, a mutational path is considered accessible if each mutation in the path increases binding affinity monotonically [88]. Figure 2.2c shows the fraction of accessible paths to the highest-affinity binding site in a global peak (that is, peak accessibility), in relation to the length of the mutational path, for the empirical data and both null models. In all three cases, the fraction of accessible mutational paths to the highest-affinity site decreases with the length of the path. However, the rate of decrease for the empirical data is intermediate to that of the additive and shuffled models, indicating that the empirical landscapes are always less navigable than the additive landscapes, but more navigable than the shuffled landscapes. Even for the longest mutational paths to the highest-affinity

site, more than 20% of the paths are accessible in the empirical landscapes. Moreover, even when unbound sequences are included in the landscapes—a modification that decreases landscape navigability—the global peak remains accessible for all but the longest mutational paths (Supplementary Fig. 2.11 and Supplementary Section 2.5.5).

Taken together, these three measures of landscape navigability—number of peaks, epistasis and peak accessibility—indicate that transcription factor binding affinity landscapes are more navigable than shuffled landscapes, but less navigable than additive landscapes. This conclusion is robust to broadly varying parameter choices (δ and τ) and modelling assumptions (Supplementary Section 2.5.6 and Supplementary Figs 2.12– 2.30). Although experiments are required to determine which parameters and assumptions best reflect the true binding affinity landscapes, the following *in vivo* analyses suggest that the baseline parameter combination studied here provides meaningful information about TF binding in both yeast and mouse, two highly diverged eukaryotic species.

2.2.5 Navigability influences the *in vivo* abundance of binding sites

Landscape navigability varies among TFs within the same species. This led us to reason that global peak sequences from more navigable landscapes might be more abundant in the regulatory regions of living organisms than global peak sequences from less navigable landscapes. The reason is that smooth landscapes pose fewer obstacles to the evolution of global peak sequences than rugged landscapes. To test this hypothesis, we consider two sources of *in vivo* data from 14 cell and tissue types in *Mus musculus*: RNA sequencing (RNA-seq) transcript abundance estimates [325] and maps of genome-wide DNase I footprints [326]. The RNA-seq data indicate which TFs are expressed in each of the 14 cell and tissue types, which is important because we only expect landscape navigability to impact the *in vivo* abundance of a binding site if its cognate TF is expressed in that cell or tissue type. The DNase I footprints demarcate DNA sequences in open chromatin that are bound by protein [327] ('Methods') and can therefore be used to predict TF binding

sites.

For the 187 murine TFs in our dataset that are expressed in a given cell or tissue type ('Methods'), we determine the *in vivo* abundance of the TF's highest-affinity binding site by counting the number of times the site appears in DNase I footprints that are predicted to bind the TF. We determine the statistical significance of each count by comparing it with the number of times the sequence is expected to appear in stretches of DNA that have the same length and the same mono- and di-nucleotide frequencies as the footprints ('Methods'). We find that the highest-affinity sites in landscapes with multiple peaks are less abundant in regulatory regions genome-wide than those from landscapes with a single peak across all of the 14 cell and tissue types (Supplementary Fig. 2.31), as shown for heart tissue in Fig. 2.3a (Wilcoxon rank-sum test, P value $= 1.42 \times 10^{-6}$). We also observe that highly accessible global peak sequences are more abundant in protein-bound regions of the mouse genome than less accessible global peak sequences across 10 of the 14 cell and tissue types (Fig. 2.3b, Supplementary Fig. 2.32). As a negative control, we repeat the above analyses using the DNase I hypersensitive regions that flank the footprints, rather than the footprints themselves. For this control data, the abundance of the highest-affinity sites is not significantly associated with the number of peaks in any of the 14 cell and tissue types. Also, for the control data and in 13 of the 14 tissue types, the abundance of the highest-affinity sites is not significantly associated with peak accessibility; in the remaining tissue (heart), this association is only marginally significant (Spearman's rank correlation coefficient $= 0.27$, P value $= 0.034$). Importantly, the effects of landscape navigability on binding site abundance still hold after controlling for binding affinity (Supplementary Section 2.5.7), although statistical significance is lost in three tissues for peak accessibility. Moreover, our observations hold in all cell and tissue types after controlling for the information content of each TF's PWM (Supplementary Section 2.5.7). Taken together, these findings suggest that landscape navigability has influenced the evolution of TF binding sites in the mouse genome.

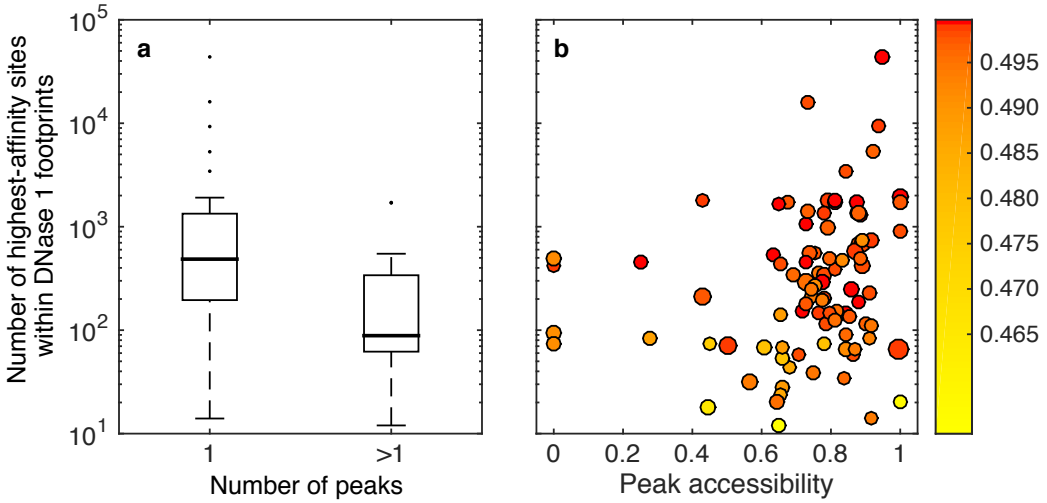


Figure 2.3: *In vivo* binding site abundance correlates with landscape navigability. **a**,**b**, The vertical axis of each panel indicates the abundance of a TF's highest-affinity site in protein-bound regions of the *M. musculus* genome in heart tissue (according to DNase I footprint data; 'Methods'). In panel **a**, the horizontal axis indicates the number of peaks and classifies landscapes into single-peaked and multi-peaked categories. Global peak sequences from single-peaked landscapes are more abundant than those from multi-peaked landscapes (Wilcoxon rank-sum test, P value $= 1.42 \times 10^{-6}$). The thick horizontal line in the middle of each box represents the median of the data, while the bottom and top of each box represent the 25th and 75th percentiles, respectively. In panel **b**, the horizontal axis shows peak accessibility through mutational paths of length four (Spearman's rank correlation coefficient $= 0.27$, P value $= 9.2 \times 10^{-3}$), which are the most abundant paths in our dataset (Fig. 2.2c). Each circle corresponds to a single TF expressed in heart tissue. Circle colour indicates the binding affinity of the TF's highest-affinity site (darker = higher; colour bar). Circle size corresponds to the TF's expression level (larger = higher, 'Methods'). Note the logarithmic scale of the y axis for both panels.

2.2.6 Gene expression reflects landscape topography

Gene expression levels can be fine-tuned via affinity-altering mutations in TF binding sites [159, 190]. Models of regulatory evolution commonly assume a direct mapping between binding affinity and gene expression [328, 329], such that monotonic changes in binding affinity lead to monotonic changes in gene expression. We test this assumption with *in vivo* gene expression data from a recent high-throughput promoter screen in *Saccharomyces cerevisiae* [190]. These data comprise replicated *in vivo* gene expression measurements for every single-base-pair and many double- and triple-base-pair mutants of a TF's con-

sensus binding site ('Methods'). PBM data are available for two of these TFs (Gcn4 and Fhl1), facilitating the superposition of *in vivo* transcriptional output with *in vitro* binding affinity. For the subset of each TF's genotype network where *in vivo* expression data are available, we determine whether gene expression levels increase monotonically along accessible mutational paths to the site with the highest affinity. For Gcn4, all 71 accessible mutational paths exhibit monotonic increases in gene expression (P value = 0.01, permutation test; Fig. 2.4a) and for Fhl1 all except one of the 37 accessible mutational paths exhibit monotonic increases in gene expression (P value = 0.001, permutation test; Fig. 2.4b). These findings suggest that, at least for these two TFs, the navigability of the adaptive landscape of binding affinity facilitates the evolution of increased gene expression.

We also compare the binding affinity landscapes of Gcn4 and Fhl1 to incomplete landscapes constructed using gene expression data from the high-throughput promoter screens. To do this, we develop a measure of similarity between the complete *in vitro* landscapes and the incomplete *in vivo* landscapes. This measure is simply the sum of the absolute differences between binding affinity and gene expression for each of the binding sites. We reason that if the *in vitro* binding affinity landscapes are truly informative of *in vivo* gene expression, then this sum will be significantly smaller than expected, if instead the gene expression measurements are randomly permuted. To test this hypothesis, we compare the observed sum to a null distribution of sums obtained via 10^5 random permutations of the gene expression data. The fraction of permuted datasets in which the sum is smaller than that of the measured expression data yields an empirical P value for each TF. Based on this test, *in vitro* binding affinity is indeed informative of *in vivo* gene expression (Gcn4: P value = 0.0011; Fhl1: P value = 0.0125; Fig. 2.4c,d), which provides additional validation of the binding affinity landscapes studied here.

2.2.7 Global peak breadth affects the diversity of binding sites

We have shown that global peaks typically comprise many different binding sites of similar affinity. The broader a peak is, the more sequences it has that have mutant neighbours

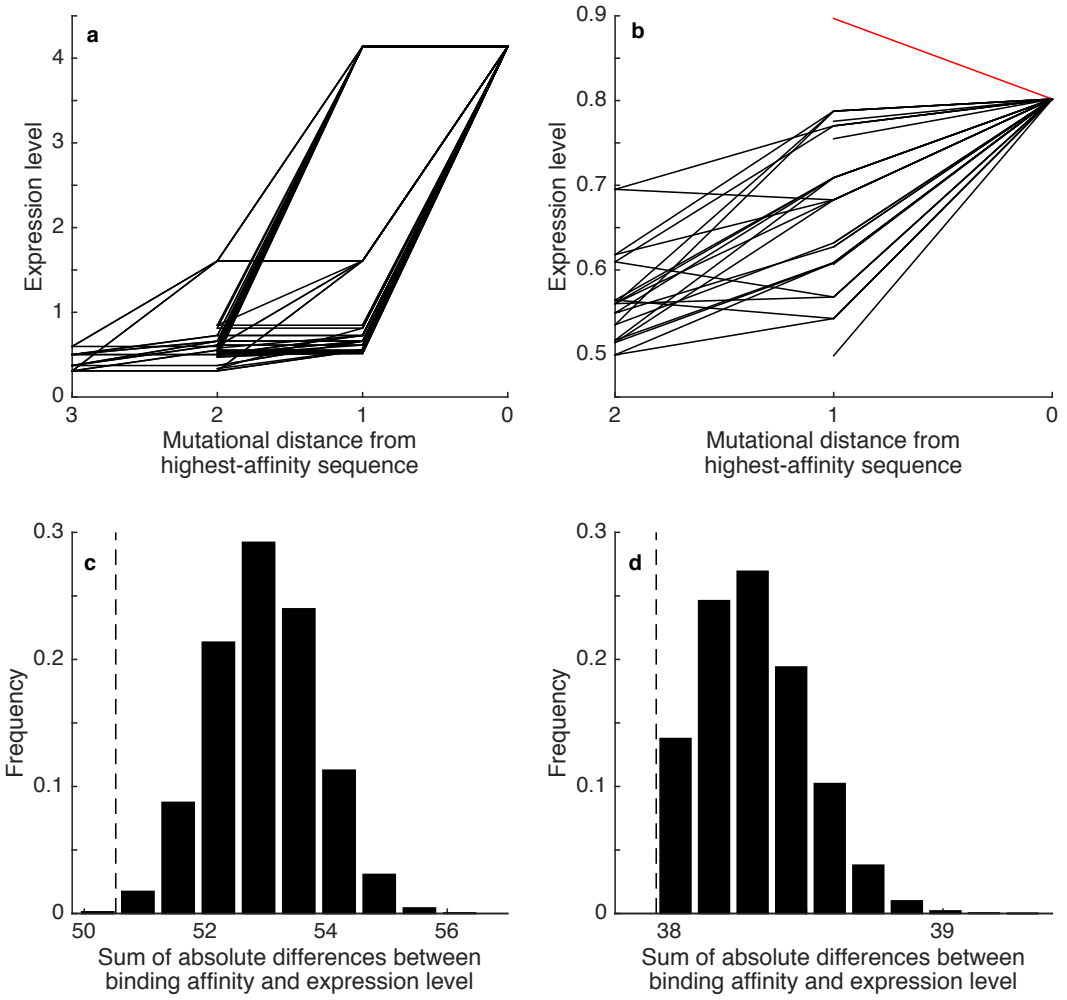


Figure 2.4: Gene expression increases along accessible mutational paths and reflects landscape topography. **a**,**b**, Each line shows the expression levels of the sequences in an accessible mutational path to the highest-affinity sequence for the yeast TFs Gcn4 (a) and Fhl1 (**b**). Black lines denote accessible mutational paths in which gene expression increases monotonically. The red line denotes the single accessible mutational path in which gene expression does not increase monotonically. Note that although expression appears to decrease along some of the black lines in panel **b**, it does not decrease beyond the noise threshold (δ) necessary to classify the trend as a true expression decrease ($\delta = 0.7$ for Gcn4 and $\delta = 0.1$ for Fhl1; 'Methods'). Note also that the highest-affinity sequence does not necessarily correspond to the sequence with the highest level of expression (red line in panel **b**). **c**,**d**, Vertical dashed lines indicate the observed sum of the absolute differences between binding affinity and gene expression, and the black bars show the null distribution of this sum for 10^5 random permutations of the gene expression values for the TFs Gcn4 (**c**) and Fhl1 (**d**).

with approximately the same affinity. One would thus expect that sequences in broader peaks could evolve more freely by means of nucleotide changes and thus accumulate

greater genetic diversity. In contrast, mutations in the binding sites of narrower peaks—those with fewer binding sites—will more often lead to a decrease in binding affinity and thus be eliminated. To find out if global peak breadth has any effect on binding site evolution, we analyse whether global peak breadth affects genetic diversity in binding sites. Using single nucleotide polymorphism data across 19 strains of *S. cerevisiae* [330], we calculate the diversity within binding sites [331] of 23 different TFs as the average Shannon diversity index per site ('Methods'). Indeed, the broader the global peak of a TF's affinity landscape, the greater the average diversity of polymorphic sites that bind the TF (Spearman's rank correlation coefficient = 0.61, P value = 0.002; Fig. 2.5). We emphasize that this trend is not driven by TF specificity, since the information content of the TFs' PWMs exhibit no correlation with binding site diversity (Spearman's rank correlation coefficient = –0.08, P value = 0.71). This finding provides further validation that the topography of a binding affinity landscape can impact the evolution of TF binding sites.

2.3 Discussion

We have used measurements from PBMs to construct and analyse more than 1,000 complete, empirical adaptive landscapes, each describing the binding affinity of a TF to all possible short DNA sequences. Such landscapes are important objects of study, because changes in the level, location or timing of gene expression commonly underlie evolutionary innovations [332, 333] and gene expression patterns are readily fine-tuned via small changes in binding affinity [159, 190]. Understanding how an evolutionary process might navigate a TF binding affinity landscape is therefore an important step towards understanding how gene regulatory programs evolve. Here, we have taken this step, demonstrating that the navigability of binding affinity landscapes is intermediate to that of additive and shuffled (rugged) landscapes, but closer to the additive expectation, in terms of the number of peaks and the incidence of three forms of epistasis.

Our measures of landscape navigability allow us to understand how individual binding sites can evolve towards a higher binding affinity in the absence of confounding factors

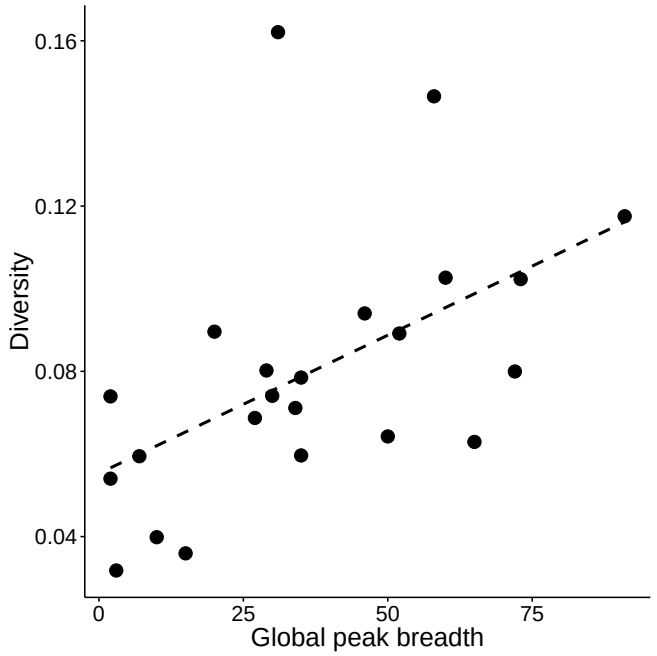


Figure 2.5: Global peak breadth influences the diversity of TF binding sites in the yeast genome. Scatter-plot for 23 TFs showing the relationship between the number of binding sites in the global peak (that is, global peak breadth) of a particular TF and the average diversity of all its polymorphic binding sites in 19 strains of *S. cerevisiae* ('Methods'; Spearman's rank correlation coefficient = 0.61, P value = 0.002). The dashed line represents the best linear-regression fit to the data.

(such as chromatin context [334]), a pursuit that is motivated by earlier theoretical work [317, 319, 328] and by several empirical observations, which indicate that high-affinity sites are used preferentially *in vivo*. For example, studies of blastoderm patterning in *Drosophila melanogaster* have shown that high-affinity sites typically reside near probable functional targets, whereas low-affinity sites are more often found near genes that are not transcribed in the early embryo [335] and that are less likely to drive expression in transgenic reporter assays [336]. Moreover, in both microbes and humans, affinity-decreasing mutations are predominately under negative selection, whereas affinity-increasing mutations are under positive selection [337, 338], consistent with the inferred monotonic increases in organismal fitness that accompany increased binding affinity [319, 320, 339]. Nevertheless, low-affinity sites do sometimes play important regulatory roles [340–343] and it is therefore worth noting that, by symmetry, accessible paths to the highest-affinity site

are also accessible in the opposite direction by permitting only monotonically decreasing changes in binding affinity. More generally, the landscapes constructed here can easily be transformed to study selection for low or intermediate binding affinity, by assuming that the fitness conferred by a binding site is a decreasing function of the difference between the site's affinity and an arbitrary optimal affinity [328]. Understanding how such transformations affect the navigability of TF binding affinity landscapes is an exciting direction for future work.

To summarize, while our analyses of *in vitro* and *in vivo* measurements of TF–DNA interactions have some caveats (Supplementary Section 2.6), they suggest that the navigability of TF binding affinity landscapes has left a trace in the portfolios of regulatory DNA within the mouse and yeast genomes. Landscape navigability may therefore have contributed to the enormous success of transcriptional regulation as an evolutionary mechanism for generating variation and innovation.

2.4 Materials and methods

2.4.1 *In vitro* data

The *in vitro* data we studied came from PBMs [184, 197], which measure the binding affinity of a TF to all 32,896 possible eight-nucleotide, double-stranded DNA sequences. There were $(4^8 - 4^4)/2 + 4^4 = 32{,}896$ sequences, rather than $4^8 = 65{,}536$ sequences, because each sequence was merged with its reverse complement and because there were 4^4 sequences that were identical to their reverse complement and therefore could not be merged.

We had three criteria for including a TF in our dataset. First, it had to be analysed on two different PBM designs. Second, it had to bind at least one DNA sequence with an E-score above 0.45, as this indicates a high level of data quality [196]. Third, its genotype network (Fig. 2.1a) had to contain at least one square to permit the analysis of epistasis (Box 1). Based on these criteria, we obtained PBM data for 42 yeast [194] and 104 mouse TFs [195] from the UniPROBE database [322] and 991 TFs belonging to 129 different

eukaryotic species (including 4 additional TFs for *S. cerevisiae* and 83 for *M. musculus*) from the CIS-BP database [196]. In total, our dataset comprised 1,137 TFs, representing 129 eukaryotic species and 62 DNA-binding domain structural classes.

For each TF, the PBM data included a non-parametric, rank-based E-score for each of the 32,896 DNA sequences. The E-score is a variant of the Wilcoxon–Mann–Whitney statistic [184] and ranges from –0.5 (most disfavoured site) to 0.5 (most favoured site). E-scores correlate with the relative dissociation constants of TFs [184, 195] and can be used as a proxy for relative binding affinity. We therefore refer to this measure as binding affinity and used it to delineate bound from unbound sequences and as the quantitative phenotype that defines the surface of our adaptive landscapes. Except for the 42 yeast TFs from Zhu *et al.* [194], the data also included a median signal intensity Z-score for each site. We used this score as an alternative proxy for binding affinity in Supplementary Section 2.5.6.5. Following earlier work [156, 194], we only considered a sequence as bound by a TF if its E-score exceeded a threshold of 0.35. The reason for this threshold was that it has precedent [195, 323] and, more importantly, an analysis of the relationship between the E-score and false discovery rate (FDR) in 104 mouse TFs [195] revealed that all sequences with an E-score exceeding 0.35 had an FDR below 0.001. This threshold could therefore be used to delineate sequences that are specifically bound by a TF from those that are unspecifically bound. In addition to the threshold $\tau = 0.35$, we conducted a sensitivity analysis of our results by broadly varying τ (Supplementary Section 2.5.6.2).

2.4.2 *In vivo* data

We collected DNase I footprints for 14 cell and tissue types in *M. musculus* [326]. These genome-wide data specify DNA sequences that are in open chromatin and bound by protein, at single nucleotide resolution. For each of the 187 murine TFs in our dataset, we used FIMO [344] and the TF's PWM (obtained from UniPROBE [322] and CIS-BP [196]) to scan these footprints for potential binding sites using $P < 1 \times 10^{-4}$ as a threshold. We then counted the number of times that each eight-nucleotide DNA sequence appeared in the predicted binding sites, considering both strands of the DNA. To determine the

statistical significance of each count, we compared it with the number of times the same sequence is expected to appear in stretches of DNA that have the same length and mono- and di-nucleotide frequencies as the footprints of each cell and tissue type, thus controlling for the GC content of the footprints. Following van Helden *et al.* [345], the statistical significance of each observed count was determined using the binomial formula, with a conservative significance threshold of $P < 1 / 32{,}896$ (the inverse of the number of possible eight-nucloetide DNA sequences). Across the 14 cell and tissue types, the counts for the highest-affinity sites of each of the 187 mouse TFs in our dataset ranged from 0 to 73,174.

We also collected the DNase I hypersensitive regions that flank the footprints in each of the 14 cell and tissue types. We used these regions to perform a negative control, in which we correlated our landscape navigability measures with *in vivo* binding site abundance within regions of open chromatin that do not show evidence of protein binding. Specifically, we counted the number of times that each eight-nucleotide DNA sequence appeared in the DNase I hypersensitive regions, after having removed the footprints of the 14 cell and tissue types. To determine the statistical significance of each count, we compared it with the number of times the same sequence is expected to appear in stretches of DNA that have the same length and mono- and di-nucleotide frequencies as the DNase I hypersensitive regions of each cell and tissue type, again using the binomial formula and a stringent significance threshold ($P < 1 / 32{,}896$).

To determine which of the 187 TFs were expressed in each of the 14 cell and tissue types, we collected RNA-seq data for the same cell and tissue types [325] and used cufflinks [346] to calculate the FPKM (fragments per kilobase of transcript per million fragments mapped) for each TF. We considered a TF as expressed in a given cell or tissue type if its FPKM $\geqslant 1$. The number of TFs expressed in a given cell or tissue type ranged from 89 to 133.

For two yeast TFs (Gcn4 and Fhl1), we also collected gene expression data from high-throughput promoter screens [190]. These data include gene expression measurements from a large library of engineered promoters, each 150 nucleotides long. Each of these

promoters contains between zero and three point mutations to the TF's consensus sequence. For Gcn4, which binds a seven-nucleotide sequence, this library includes the consensus sequence (TGACTCA), all 21 single mutants, 42 double mutants and 10 triple mutants (73 sequences in total). For Fhl1, which binds an eight-nucleotide sequence, this library includes the consensus sequence (GACGCAAA), all 24 single mutants, 56 double mutants and 10 triple mutants (90 sequences in total). For each promoter in the library, the data include gene expression measurements from two biological replicates. We used the average gene expression measurement per promoter.

To map the gene expression measurements for the seven-nucleotide sequences bound by Gcn4 onto the eight-nucleotide sequences for which we had PBM data, we first located each of the 73 seven-nucleotide sequences within the 150-nucleotide promoters. We then padded each of the seven-nucleotide sequences with one nucleotide upstream and one nucleotide downstream from its respective promoter, forming two eight-nucleotide sequences. We assigned the gene expression measurement for the seven-nucleotide sequence to each of the eight-nucleotide sequences that were formed in this fashion. This procedure generated gene expression measurements for 132 eight-nucleotide sequences.

For the subset of the sequences for which we had measurements of both *in vitro* binding affinity and *in vivo* gene expression, we determined the number of accessible mutational paths to the sequence with the highest level of binding affinity. For Gcn4, there were 71 accessible mutational paths: 30 of length 3, 37 of length 2 and 14 of length 1. For Fhl1, there were 37 accessible mutational paths: 22 of length 2 and 15 of length 1. For each TF, we then determined the fraction of these accessible paths in which gene expression increased monotonically. To determine whether two gene expression levels truly differed from one another, we used a noise threshold that covered the same proportional range of expression levels as covered by the noise threshold δ in the range of affinity values. To calculate the statistical significance of the fraction of accessible paths in which gene expression increased monotonically, we performed a permutation test, in which we randomly permuted the gene expression measurements of the sequences, while preserving

their binding affinities. We repeated this process 1,000 times for both Gcn4 and Fhl1.

We obtained the genomic coordinates of TF binding sites in *S. cerevisiae* from a map of conserved regulatory sites [329] and analysed binding sites detected with a stringent binding P value cut-off of 0.001, but no conservation cut-off. We obtained single nucleotide polymorphism (SNP) data for 19 different strains from the Saccharomyces Genome Resequencing Project (SGRP) [330]. The TF binding sites were based on the January 2006 Saccharomyces Genome Database (SGD) version of the reference strain *S. cerevisiae* S288c genome sequence, while the SNP data were based on the January 2010 version. We used liftOver to convert both sets of genomic coordinates to the coordinates in the February 2011 version. Then we used the *intersect* function from the BEDTools suite (version 2.25.0) [347] to determine the presence of SNPs within binding sites. Using these data, we calculated the genetic diversity of polymorphic TF binding sites across the 19 yeast strains. For each position j in a binding site we calculated the Shannon diversity index (H):

$$H_j = -\sum_i p_i \log_2(p_i), \tag{2.1}$$

where p_i is the frequency of allele i, which we computed as the fraction of strains with allele i. We computed the diversity (D) of a binding site as the average of H over all L positions:

$$D = \frac{1}{L}\sum_{j=1}^{L} H_j, \tag{2.2}$$

2.4.3 Genotype networks

The procedure for constructing genotype networks of TF binding sites has been described elsewhere [156]. In brief, for each TF, we first determined the set of sequences that were bound by the TF (E-score > 0.35). We then used an alignment algorithm to calculate the mutational distance between all pairs of bound sequences. Finally, we used these mutational distances to define the edges of the genotype network, connecting two sequences

if they differed by a single small mutation. The mutations we considered were point mutations and small indels that shift an entire contiguous binding site by a single base [156].

2.4.4 Quantitative measures of landscape navigability

We used several measures to quantify the navigability of an adaptive landscape. All of them were parameterized by δ, a threshold value that is used to determine whether two affinity values truly differ from one another. This parameter is necessary because PBM data are inherently noisy [197]. For each TF, we calculated δ as the residual standard error of a linear regression between the affinity values of all bound sequences from the two replicate PBMs. Thus, each TF had its own δ, which reflected the noise in the replicated PBM measurements for that particular TF. We considered that the binding affinity E_i of site i was greater than the binding affinity E_j of site j if $E_i > E_j + \delta$. Analogously, we considered that the binding affinity E_i of site i was less than the binding affinity E_j of site j if $E_i + \delta < E_j$. Otherwise, we considered that we could not differentiate between the two affinity values. The average empirical value of δ across the 1,137 TFs was 0.028, which covered 18.4% of the range of affinity values for bound sequences ($0.35 < E$-score ≤ 0.50). Our criterion for considering two affinity values as different was therefore highly conservative. In addition to these empirical values of δ, we conducted a sensitivity analysis of our results by broadly varying δ (0, 0.001, 0.01, 0.03 and 0.05) above and below the average empirical value (Supplementary Section 2.5.6.1).

Our first measure of the navigability of an adaptive landscape was its number of peaks. To detect a peak, we followed a procedure similar to one previously described [79]. We categorized each sequence in a genotype network as belonging to a peak (either as an individual sequence or as a member of a plateau) or not. To do this, we selected sequences in decreasing order of binding affinity to seed a breadth-first search of the genotype network. Each iteration of the search considered sequences that were an additional mutational step away from the seed sequence. In the first iteration of the search, we determined whether the seed sequence was a peak or not; it was considered a peak if all of its neighbours in

the genotype network had a lower binding affinity and was not a peak if at least one of its neighbours had higher binding affinity. If the seed sequence had at least one neighbour with an affinity that was neither greater nor lower, then these neighbours were retained as belonging to a plateau that may be a peak and the breadth-first search continued. If at any subsequent iteration of the search a sequence was found that neighboured any of those on the plateau and had an affinity that was higher than the seed sequence, then the seed sequence did not belong to a peak and the search was halted. If at any iteration of the search, all sequences that neighboured those on the plateau had a lower affinity than the seed sequence, then the seed sequence belonged to a plateau that was a peak.

Our second measure of landscape navigability was based on the concept of accessible mutational paths [82, 88, 348]. A mutational path is the shortest path that connects two bound sequences, i and j, on a genotype network, such that sequence i can be transformed into sequence j via a series of intermediates that are also on the genotype network. A mutational path from sequence i to sequence j is considered accessible if binding affinity increases monotonically along the path. We report the fraction of mutational paths that are accessible, starting from all sequences in the genotype network and ending at the highest-affinity site in the global peak. We refer to this fraction as peak accessibility. We note that, by symmetry, accessible paths to the highest-affinity site are also accessible in the opposite direction by permitting only monotonically decreasing changes in binding affinity. Accessibility of low-affinity binding sites is important, as such sites are known to play crucial roles in the regulation of certain genes [341, 343, 349].

Our final measures of landscape navigability pertained to epistasis, motivated by a recent debate over the significance of non-additive interactions between the individual bases of TF binding sites in their contribution to binding affinity [201, 202]. Epistasis can have detrimental effects on landscape navigability. We quantified the incidence of three classes of epistasis for each of the 1,137 adaptive landscapes [82]. First, we detected all squares in each TF's genotype network. A square is a quadruplet of sequences that contain a binding site, two of its one-mutant neighbours, and the double mutant that can

be formed from the single mutants (Box 1). Second, we designated the highest-affinity sequence as the 'double mutant', thus forcing the labelling of the other three sites as the wild type or single mutants. Third, we calculated the magnitude of epistasis as:

$$\epsilon = E_{AB} + E_{ab} - E_{Ab} - E_{aB}, \tag{2.3}$$

where E_AB is the binding affinity of the double mutant, E_ab is the binding affinity of the wild type and E_Ab and E_aB are the binding affinities of the single mutants. We only considered a mutational pair to be epistatic if $|\epsilon|$ was greater than or equal to the noise threshold δ. If this condition was met, we classified the epistatic interaction as magnitude epistasis, simple sign epistasis or reciprocal sign epistasis [149]. The interaction was classified as magnitude epistasis when the following relation held:

$$|\Delta E_{ab\to Ab} + \Delta E_{aB\to AB}| = |\Delta E_{ab\to Ab}| + |\Delta E_{aB\to AB}|, \tag{2.4}$$

where ΔE is the 'mutational effect', that is, the change in binding affinity caused by a mutation (for example, $ab \to Ab$). The interaction was classified as simple sign epistasis when the following relation held:

$$|\Delta E_{ab\to Ab} + \Delta E_{aB\to AB}| < |\Delta E_{ab\to Ab}| + |\Delta E_{aB\to AB}|, \tag{2.5}$$

The interaction was classified as reciprocal sign epistasis when both equation (2.3) and the following relation held:

$$|\Delta E_{ab\to aB} + \Delta E_{Ab\to AB}| < |\Delta E_{ab\to aB}| + |\Delta E_{Ab\to AB}|, \tag{2.6}$$

If a mutational effect was smaller than δ, we assigned it a value of zero. If all mutational effects were smaller than δ, even despite $|\epsilon| \geq \delta$ being true, then we classified the interaction as non-epistatic. Taken together with the fact that we excluded unbound sequences from these calculations, our measures of epistasis were conservative (Supplementary Section 2.5.8 and Supplementary Fig. 2.33).

2.4.5 Null models

For each of the 1,137 TFs, we considered two null models. Both changed the topography of the adaptive landscape, but maintained the structure of the underlying genotype network. The additive model was based on the PWM of a TF. A PWM represents the binding preferences of a TF as a $4 \times L$ matrix, where each row corresponds to one of the four bases and each column corresponds to one of the L positions in the binding site. Each matrix entry, $f_{i,b}$, is the frequency of base b at position i. We obtained the PWMs of the 1,137 TFs from the UniPROBE [322] and CIS-BP [196] databases.

Since the width L of a TF's PWM may not equal eight, we used a sliding-window to assign a score (S_{pwm}) to each of the eight-nucleotide sequences in each genotype network. Specifically, we slid each sequence in a genotype network through the corresponding TF's PWM from left to right, assigning a score to each of the subsequences. The sliding-window procedure was carried out in such a way that the first scored subsequence was the single right-most position of the sliding sequence and occupied the left-most column of the PWM. The procedure ended with a subsequence that corresponded to the single left-most position in the sequence and occupied the right-most column of the PWM. We then repeated this process with the sequence?s reverse complement. We took the maximum of these scores as the score for the sequence. As an example, if L was equal to eight, then we took the maximum of 30 separate scores: 15 for the sequence and another 15 for its reverse complement. Each of these scores was calculated as

$$S_{pwm} = \sum_{i=1}^{l} f_{i,b} I(i), \tag{2.7}$$

where l is the window length of the sliding sequence and $I(i)$ is the information content at position i:

$$I(i) = 2 + \sum_{b} f_{i,b} \log_2 f_{i,b}, \tag{2.8}$$

Base matches at positions with high information content thus contribute more to a

sequence's overall score than base matches at positions with low information content. Importantly, this scoring technique was purely additive, that is, the contribution of each binding site position to the overall score was independent of the other positions in the binding site. To facilitate comparison among TFs, we normalized the scores by the maximum score in each genotype network. To analyse the navigability of landscapes constructed with this model, we used a noise threshold δ that covers the same range of scores as the δ used for the empirical data.

In the shuffled model, we randomly permuted the binding affinities of the sequences in the genotype network, yielding a rugged landscape topography. Since this process was stochastic, we repeated it 1,000 times per TF. In these landscapes, we used the same δ as that used for the empirical landscapes, because, unlike the additive model, the shuffled null model does not change the range of affinity values.

2.5 Supplementary results

2.5.1 Summary statistics of genotype networks

Here, we summarize some of the structural properties of the genotype networks that serve as the substrate of the adaptive landscapes we study. We observe that depending on the TF, between 9 and 1,186 sites are bound. For 99.6% of the TFs (1,132 of 1,137), the majority of bound sites form a single connected component in this network, which we refer to as the dominant network, and for 53% of the TFs (600 of 1,137, Supplementary Fig. 2.1A), all of the bound sites belong to this dominant network (corroborating our previous findings based on fewer TFs [156]). For the vast majority of TFs, the vast majority of bound sequences are in the dominant genotype network (Supplementary Fig. 2.2), while many of the non-dominant networks comprise very few sequences, and predominantly only a single sequence (Supplementary Fig. 2.3). Therefore, we carry out all of our analyses on the dominant genotype networks, which contain on average 392 sites but vary broadly in size among the TFs (Supplementary Fig. 2.1B).

2.5.2 Global peaks are usually organized into broad plateaus

We find that global peaks in binding affinity landscapes rarely comprise only a single sequence; rather they are plateaus typically made up of dozens to hundreds of sequences. The average plateau is large and comprises 10.4% of all binding sites in the landscape (Supplementary Fig. 2.7), many more than expected in the shuffled model, but far fewer than expected in the additive model. While larger landscapes also tend to have larger plateaus (Spearman's rank correlation coefficient = 0.7, P value $< 2.2 \times 10^{-16}$), the fraction of a landscape's binding sites falling on a plateau decreases with landscape size (Supplementary Fig. 2.6B; Spearman's rank correlation coefficient = -0.35, P value $< 2.2 \times 10^{-16}$).

This observation of global peaks being organized into broad plateaus is in line with genome-wide assays of TF binding, which often find non-consensus sequences in bound regions of the genome [335, 350]. We stress, however, that our measure of peak breadth is sensitive to the use of E-scores (rather than Z- scores) as the quantitative phenotype in our landscapes (Supplementary section 2.5.6.5), and this sensitivity should be taken into account when interpreting our findings.

2.5.3 Why epistasis occasionally appears in the additive null model

Epistasis appears at very low frequencies in the additive model. For example, an average of 0.036% of squares per genotype network exhibit simple sign epistasis, and 0.007% exhibit reciprocal sign epistasis. On the rare occasion where epistasis appears in the additive model, it stems directly from our sliding-window approach for scoring sequences ('Methods'). Supplementary Fig. 2.8 shows an example. The mouse TF Arid5a has a PWM that is 14 nucleotides wide (Supplementary Fig. 2.8A,B). The mutational pair AATTTTAA and AAATATAA exhibits reciprocal sign epistasis (Supplementary Fig. 2.8C), since both mutational intermediates have lower binding affinity. This results from the fact that the highest score for the sequence AATTTTAA occurs when it is aligned to positions 3

through 10 in the PWM, whereas the highest score for the other three sequences in the square occur when they are aligned to positions 2 through 9. If AATTTTAA were also aligned to positions 2 through 9, then this square would not exhibit epistasis.

While our sliding-window approach is therefore responsible for a very low incidence of epistasis in the additive null model, we believe that it is the most sensible approach for scoring sequences, for at least two reasons. First, most of the PWMs in our dataset are not eight nucleotides wide, and in these cases it is simply necessary to use a sliding window. Second, even for the PWMs that are eight nucleotides wide, the sliding window is superior to a fixed window. To understand why, consider a hypothetical example where the information content in each position of the PWM is maximal, i.e., there is no variation in the nucleotide frequencies per position. Such a PWM could be represented by a single consensus sequence, say TATATATA. Using a fixed window, the sequence ATATATAT would not match the PWM in a single position and the sequence would receive a score of zero. However, by sliding the sequence by only one position, it will receive a high score, as it should, since seven of its eight nucleotides exactly match the consensus.

2.5.4 Sign epistasis preferentially occurs among nucleotides that are near one another in the binding site

The prevalence of epistatic interactions in the TF binding sites of 411 mouse and human TFs was recently investigated using data from high-throughput SELEX [324]. For each TF, pairwise epistasis was estimated as the ratio of the observed frequency of a mutational pair in a binding site to its expected frequency, which was calculated using the TF's PWM. This ratio therefore describes a deviation from additivity. Based on this analysis, the authors concluded that the individual nucleotides of a binding site generally contribute additively to binding affinity, and that when epistatic interactions do occur, they occur preferentially among nucleotides that are near one another in the binding site.

Our analyses generally agree with these observations, but we also extend these earlier findings by partitioning epistasis into three distinct classes that have increasingly detri-

mental effects on landscape navigability: magnitude epistasis, simple sign epistasis, and reciprocal sign epistasis (Box 1). Specifically, we find that both simple sign epistasis (Supplementary Fig. 2.9B) and reciprocal sign epistasis (Supplementary Fig. 2.9C) decrease as the distance between the binding site positions increases, but we observe no such trend for magnitude epistasis (Supplementary Fig. 2.9A). Thus, the two classes of epistasis that hamper landscape navigability the most—simple and reciprocal sign epistasis—occur preferentially among nearby binding site positions, whereas the class of epistasis that does not affect landscape navigability—magnitude epistasis—has no such preference.

2.5.5 Peak accessibility decreases when unbound sequences are included

As with our measure of epistasis (Supplementary section 2.5.8), we restricted our measure of peak accessibility to bound sequences. Including unbound sequences can only decrease the navigability of a landscape, and therefore decrease peak accessibility. To determine the extent of this decrease, we repeated our peak accessibility analyses for mutational paths that include unbound sequences. Supplementary Fig. 2.11 shows that the measures of peak accessibility reported in the main text are always higher than those that include unbound sequences (especially at longer mutational distances), as expected. The reason is twofold. First, since mutational paths always start at a bound sequence (i.e., $\tau > 0.35$), any path that includes an unbound sequence (i.e., $\tau \leq 0.35$) is by definition inaccessible. Second, the inclusion of unbound sequences creates new shortest paths to the global peak that are inaccessible and that occasionally supplant longer, accessible shortest paths to the global peak that were present in the genotype networks studied in the main text.

2.5.6 Sensitivity analyses

Our study employs two parameters, one for determining whether two DNA sequences differ in their binding affinity (δ) and another to determine which DNA sequences are bound by a TF (τ). The threshold δ is important because it accounts for the experimental noise in PBM data [197]. Increasing δ has the effect of 'smoothing' the landscapes

(Fig. 2.1b), because larger values of δ make it less likely that two sequences have different binding affinities. We use a unique threshold δ for each TF, which is derived from noise in experimental and replicated PBM measurements for that particular TF ('Methods'). Increasing the affinity threshold τ reduces the number of bound sequences, and decreases the size of a landscape (Fig. 2.1c). Below, we assess the sensitivity of our main results to broadly varying parameter values for τ and δ. Moreover, we determine whether our main results differ among TFs from distinct DNA binding domain structural classes, as well as among TFs that bind DNA sequences shorter or longer than eight nucleotides. Finally, we show that our measure of global peak breadth is sensitive to the use of E-scores (rather than Z-scores) as a quantitative phenotype.

Specifically, we test the sensitivity of the following main results: (1) The empirical landscapes have more peaks than expected under the additive model, but far fewer than expected under the shuffled model. (2) The global peaks in the empirical landscapes comprise more sequences than expected under the shuffled model, but fewer than expected under the additive model. (3) The incidence of the three classes of epistasis in the empirical landscapes is higher than expected under the additive model, but lower than expected under the shuffled model. (4) The highest-affinity sites are more accessible in the empirical data than expected under the shuffled model, but less accessible than those expected under the additive model.

2.5.6.1 Our observations are insensitive to broadly varying thresholds for noise filtering

In the main text, we used a threshold δ to determine whether two sequences differed in their binding affinity. This threshold is important because it accounts for the inherent noise of protein-binding microarray data [197]. Here, we explore the sensitivity of our results to this parameter.

As the noise threshold δ increases, the number of peaks decreases, for the empirical data and both null models (Supplementary Fig. 2.12). This is because increasing the noise threshold makes it less likely for two sequences to have different binding affinities,

which has the effect of merging peaks. Nonetheless, even for the most permissive noise threshold of $\delta = 0.05$ (which corresponds to one-third of the range in affinity values when $\tau = 0.35$), 32% of the empirical landscapes comprise multiple peaks, as opposed to 3% of the additive landscapes and 98% of the shuffled landscapes.

Supplementary Figure 2.13 shows that the number of binding sites per global peak (i.e., global peak breadth) increases as the noise threshold δ increases, for the empirical data and both null models. This occurs because increasing δ makes the global peak more 'inclusive', such that the binding affinity of a sequence can differ from that of the highest-affinity sequence by a greater amount and the sequence will still be considered part of the global peak.

Supplementary Figure 2.14 shows that epistasis decreases as the noise threshold δ increases, for all three classes of epistasis, in the empirical data and both null models. This occurs because increasing δ decreases the likelihood that Eqs. (2.3), or (2.4) ('Methods') are satisfied, which decreases the proportion of mutational pairs that are classified as epistatic. These sensitivity analyses confirm intuition: increasing the noise threshold δ increases the navigability of an adaptive landscape. This is further evidenced by Supplementary Fig. 2.15, which shows that peak accessibility increases with δ in both the empirical data and the two null models. With the exception of magnitude epistasis—which does not affect landscape navigability—at low δ, all measurements of epistasis and peak accessibility in the empirical data are intermediate to those of the additive and shuffled null models. We therefore conclude that the adaptive landscapes of TF binding preferences are less navigable than those from the additive model, but far more navigable than landscapes from the shuffled model, and that this observation is insensitive to changes in δ.

2.5.6.2 Our observations are insensitive to broadly varying affinity thresholds for delineating bound from unbound sequences

In the main text, we considered a sequence 'bound' if its E-score exceeded 0.35. Here, we assess the sensitivity of each of our measures of landscape navigability to the choice of

binding affinity threshold τ.

As the affinity threshold τ increases, the distribution of the number of peaks shifts toward lower values, such that an increasing number of landscapes become single-peaked (Supplementary Fig. 2.16). This is true for the empirical data and for both null models, and results from the fact that increasing the binding affinity threshold decreases the number of bound sequences [156], thus pruning non-global peaks from the landscape. However, even for the most stringent threshold of $\tau = 0.45$, under which only very few sequences are considered bound [156], 14% of the empirical landscapes comprise multiple peaks, as compared to 2% of the additive landscapes and 58% of the shuffled landscapes.

Supplementary Figure 2.17 shows that the empirical distribution of the number of binding sites in the global peak (i.e., global peak breadth) is relatively unaffected by changes in the binding affinity threshold τ. This is because the sequences in these peaks typically have E-scores that exceed even our most stringent threshold of $\tau = 0.45$. In contrast, the distributions of global peak breadth under the two null models are sensitive to the binding affinity threshold, but in different ways. In the additive model, the global peak breadth decreases as τ increases. This is because the sequences at the base of each peak are gradually removed from the landscape as τ increases, analogous to the loss of exposed landmass as water levels rise. In the shuffled model, in contrast, the global peak breadth increases as τ increases, a counterintuitive observation that stems from the way these landscapes are constructed. Specifically, increasing τ decreases the number of bound sequences [156], and in a landscape with few sequences, a random permutation of binding affinities is likely to clump many of the high-affinity sequences near one another in the landscape, such that they are part of the same peak. In contrast, in a landscape with many sequences (low τ), these same high-affinity sequences are likely to be spread throughout the landscape, and thus to be part of different peaks. Despite these changes in the landscapes generated by the two null models, the global peaks in the empirical landscapes persist in comprising more sequences than expected under the shuffled model, but fewer than expected under the additive model. We therefore conclude that these

observations are insensitive to changes in τ.

Supplementary Figure 2.18 shows boxplots of the three classes of epistasis as a function of the binding affinity threshold τ, for the empirical data and the two null models. In the additive model, there is very little epistasis and it is generally insensitive to changes in τ. In the empirical data and in the shuffled model, the incidence of epistasis tends to decrease as τ increases, indicating an increase in landscape navigability. These observations are reflected in Supplementary Fig. 2.19: peak accessibility increases as the binding affinity threshold τ increases for both the empirical data and the shuffled model, but for the additive model, peak accessibility does not change with τ. Importantly, all measurements of epistasis and peak accessibility in the empirical data are intermediate to those of the additive and shuffled null models. We therefore conclude that the adaptive landscapes of TF binding affinity are less navigable than those from the additive model, but far more navigable than landscapes from a shuffled model, and that this observation is insensitive to changes in τ.

2.5.6.3 Our observations are consistent across DNA binding domains

The 1,137 TFs in our dataset represent 62 different DNA binding domain structural classes. Here, we show that our results do not vary among these classes. While we illustrate this point using data from the most prominent DNA binding domains in our dataset, the insensitivity of our results applies to the other structural classes as well (data not shown).

Supplementary Figure 2.20 shows the distribution of the number of peaks in the empirical data and the two null models for the five most prominent DNA binding domain classes. While the distributions change quantitatively, especially under the shuffled model, they do not change qualitatively. Specifically, the empirical landscapes always have more peaks than expected under the additive model, but fewer than expected under the shuffled model.

Supplementary Figure 2.21 shows the number of binding sites in the global peak (i.e., global peak breadth) in the empirical data and the two null models, for the same five DNA

binding domain structural classes. Again, the distributions do not change qualitatively: Global peak breadth is larger in the empirical landscapes than expected under the shuffled model, but smaller than expected under the additive model.

Supplementary Figure 2.22 shows that the incidence of the three classes of epistasis among the empirical data and the two null models does not vary among DNA binding domain structural classes. Epistasis in the empirical data is always intermediate to that of the additive and shuffled null models. As a consequence, peak accessibility in the empirical data is intermediate to that of the additive and shuffled null models (Supplementary Fig. 2.23). We therefore conclude that the adaptive landscapes of TF binding preferences are less navigable than those from the additive model, but far more navigable than landscapes from the shuffled model, and that this observation does not vary among the DNA binding domain structural classes considered here.

2.5.6.4 Our observations are consistent across TFs that bind shorter or longer sequences than eight nucleotides

Protein-binding microarray data characterize the binding affinity of a TF to all possible binding sites of length 8. However, many of the TFs in our dataset bind sequences that are shorter or longer than 8 nucleotides. It is therefore important to ensure that our results are insensitive to binding site length. To determine the length of a TF's binding sites, we use the TF?s position weight matrix to calculate the maximum number of contiguous nucleotides with information content (Eq. (2.8), 'Methods') exceeding 0.5 bits (bold font in Supplementary Fig. 2.8B).

We find that among the 1,137 TFs studied here, 935 bind sequences that are shorter than 8 nucleotides and 47 bind sequences that are longer than 8 nucleotides; the remaining 155 TFs bind sequences that are exactly 8 nucleotides long. Supplementary Figure 2.24 shows that the empirical landscapes comprise more peaks than expected under the additive model, but far fewer than expected under the shuffled model, regardless of binding site length. Supplementary Figure 2.25 shows that global peaks in the empirical landscapes typically comprise multiple sequences. Notably, for TFs that bind sequences shorter than

8 nucleotides, the distribution of global peak breadth shifts toward higher values under the additive model. This results from our scoring strategy, which assigns a score to each sliding subsequence, and then assigns the maximum of these scores to the sequence ('Methods'). Since many sequences will contain the same high-scoring subsequence, global peak breadth increases for TFs that bind sequences shorter than 8 nucleotides under the additive model.

Supplementary Figure 2.26 shows that the incidence of the three classes of epistasis among the empirical data and the two null models does not vary among TFs that bind sequences that are shorter or longer than 8 nucleotides. Epistasis in the empirical data is greater than expected under the additive model, but less than expected under the shuffled model. These trends are reflected in Supplementary Fig. 2.27, which depicts peak accessibility as a function of the mutational distance from the peak sequence. Peaks in the empirical adaptive landscapes are always less accessible than those from the additive model, but much more accessible than those from the shuffled model. We therefore conclude that the adaptive landscapes of TF binding affinity are less navigable than those from the additive model, but far more navigable than landscapes from the shuffled model, and that this observation is insensitive to binding site length.

2.5.6.5 Peak breadth is sensitive to the use of E-scores as a quantitative phenotype

Protein-binding microarray data typically include both an E-score and a Z-score for each DNA sequence. In the main text, we considered the E-score as the quantitative phenotype that defines the surface of each adaptive landscape. The E-score has the advantage that it is robust to outliers, insensitive to TF concentrations, and less variable across arrays. However, it has the disadvantage of compressing the dynamic range of binding affinities, particularly for E-scores near the maximum value of 0.5. Here, we assess the sensitivity of our results to the use of E-scores for the 1,095 landscape for with we have Z-scores ('Methods'), revealing that global peak breadth is sensitive to this measure, whereas our measures of landscape navigability are not.

Supplementary Figure 2.28 shows that the number of peaks in the empirical landscapes remains intermediate to that of the additive and shuffled models, although the number of peaks does increase slightly. However, the number of sequences per peak decreases significantly, relative to the landscapes constructed using E-scores. For example, the average number of sequences per peak decreases from 33.2 to 3.6, and the number of single-sequence peaks increases from 36 to 650. Thus, our finding that peaks typically comprise dozens to hundreds of sequences should be considered in the context of this sensitivity. In contrast, the three classes of epistasis that we study are insensitive to the use of E-scores (Supplementary Fig. 2.29), as are our measures of peak accessibility (Supplementary Fig. 2.30).

2.5.7 The *in vivo* relationship between landscape navigability and the abundance of binding sites is not driven by binding affinity or by information content

We have previously found that the mouse genome is enriched in high-affinity binding sites both genome-wide and within putative enhancers [156]. Therefore, we checked whether any of the *in vivo* correlations with the abundance of the highest-affinity site in DNase I footprints could be explained by a confounding relationship with binding affinity. To evaluate that possibility, we tested whether the observed differences in site abundance between single-peak and multi-peaked landscapes are still significant after adjusting for binding affinity. An analysis of covariance (ANCOVA) [351] shows that after controlling for differences in the response variable (site abundance) caused by the covariate (binding affinity), there are still significant differences caused by peak number. In both brain and MEL cells, the assumption of the homogeneity of slopes among groups—one of the underlying assumptions for the ANCOVA—is violated. However, in brain, the main effect of peak number in a linear model that incorporates the interaction between peak number and binding affinity is significant (t-test, P value = 0.038). A partial correlation analysis shows that the effect of peak accessibility still holds after controlling for binding affinity,

except for three tissue types.

To measure *in vivo* binding site abundance, we scanned DNase I footprints using FIMO [344] and each TF's position weight matrix ('Methods'). Since the information content of a TF's position weight matrix may impact the number of 'hits' detected in such genomic scans, we performed an additional ANCOVA that controlled for this potential confounding factor. However, this analysis was only necessary for brain and heart tissues since for all other cell or tissue types there was no linear relationship between site abundance and information content. This analysis shows that the effect of peak number on the abundance of the highest-affinity site in DNase I footprints remains statistically significant after controlling for the information content of each TF's position weight matrix. A partial correlation analysis shows that the same is true for peak accessibility.

Thus, we conclude that landscape navigability influences the prevalence of high-affinity TF binding sites in protein-bound regions of the mouse genome, independent of binding affinity or the information content of the TF's position weight matrix.

2.5.8 Our measures of epistasis for bound sequences are conservative

We have restricted our measures of epistasis to bound sequences (i.e. sequences with a binding affinity above τ; 'Methods') because these form the genotype networks that underlie the adaptive landscapes we study. Our measures of epistasis are therefore conservative, because they do not include unbound sequences, and therefore underestimate the level of ruggedness in a landscape. To determine how conservative our measures of epistasis are, we performed an additional analysis in which we included both bound and unbound sequences. Specifically, in each landscape, we analysed all pairs of bound sequences that differed by two mutations and used them to calculate epistasis for all possible mutational paths between these two sequences (i.e., all squares), regardless of whether the intermediate sequences were part of the genotype network or not. Supplementary Fig. 2.33 shows that the measures of epistasis reported in the main text are always lower than those that

include unbound sequences, as expected.

2.6 Supplementary discussion

Since Sewall Wright introduced the metaphor of the adaptive landscape in the early 1930s [40, 46], it has received considerable attention from theorists interested in understanding how landscape topography affects evolutionary dynamics [95]. In recent years, however, attention has shifted from theoretical landscapes to the analysis of empirical landscapes constructed from experimental data [56]. This shift has been triggered by advances in high-throughput sequencing and chip-based technologies, which have made it possible to assign phenotypes or fitness values to a large number of genotypes, thus bringing the landscape metaphor to life at unprecedented resolution. The study of empirical landscapes is currently a burgeoning area of research [26, 29–32, 56].

However, most of the empirical adaptive landscapes studied to date describe only a minute fraction of the full genotype space in which they are embedded, and they are therefore highly incomplete. For example, one of the pioneering studies of empirical adaptive landscapes considered antibiotic resistance as a function of all possible combinations of five mutations in the TEM-1 β-lactamase allele of *Escherichia coli*, and concluded that the landscape is single-peaked [88]. However, a subsequent analysis of a greater number of resistance-conferring mutations in the same allele demonstrated that this landscape is in fact multi-peaked [52], highlighting that the analysis of incomplete landscapes may produce misleading results [352].

Here, we have studied 1,137 complete adaptive landscapes of transcription factor binding affinity using experimental data from protein binding microarrays. We find that they exhibit little epistasis and contain few peaks, and these peaks are easily accessed via a series of small mutations that only move 'uphill'. We highlight that these findings go far beyond our previous work on genotype networks of transcription factor binding sites [156], in which we studied a binary phenotype—the molecular capacity of a DNA sequence to bind specific transcription factors—and ignored the quantitative measures of binding affinity that are the focus of this study. Including a quantitative measure of binding

affinity transforms 'flat' genotype networks into 'three-dimensional' landscapes.

We have characterized the incidence of epistasis as a measure of landscape navigability, motivated by a recent debate over the significance of epistatic interactions in TF binding sites [201, 202], and more generally over the prevalence of epistasis in molecular evolution [353, 354]. While such interactions have been convincingly demonstrated for some TFs [355–357], large-scale analyses of mouse and human TFs suggest that epistasis is the exception, rather than the rule, in TF binding [195, 324]. Our results extend and complement these earlier findings, by categorizing epistatic interactions as magnitude, simple sign, or reciprocal sign epistasis. This categorization reveals that both forms of sign epistasis are rare in TF binding sites, but that magnitude epistasis is common. However, there are some notable exceptions. For example, the empirical landscapes of four TFs (Dbp, bzpH, e_gw1.279.28.1, PRKRIR) exhibit the maximum value of magnitude epistasis, something that is rarely seen in the shuffled model. More generally, the empirical landscapes of 171 TFs exhibit more magnitude epistasis than the shuffled landscapes, the empirical landscapes of 25 TFs exhibit more sign epistasis than the shuffled landscapes, and the empirical landscapes of 15 TFs exhibit more reciprocal sign epistasis than the shuffled landscapes. These TFs tend to have small dominant genotype networks (the median number of bound sequences is 103), suggesting that high-specificity TFs tend to bind sequences that exhibit high levels of epistasis. Additionally, while sign epistasis preferentially occurs among bases that are near one another in the binding site—as previously reported for human TFs [324]—magnitude epistasis shows no such preference: Bases at opposite ends of a binding site are just as likely to exhibit magnitude epistasis as are adjacent bases.

Most of what was previously known about TF binding affinity landscapes came from *in vitro* evolution experiments [358] and biophysical models of TF-DNA interactions [317–320, 328, 339, 359–363]. Specifically, *in vitro* evolution experiments with the bacterial *lac* repressor have hinted that TF binding affinity landscapes are highly navigable [358], because high affinity binding sites evolve quickly from a small randomized library of

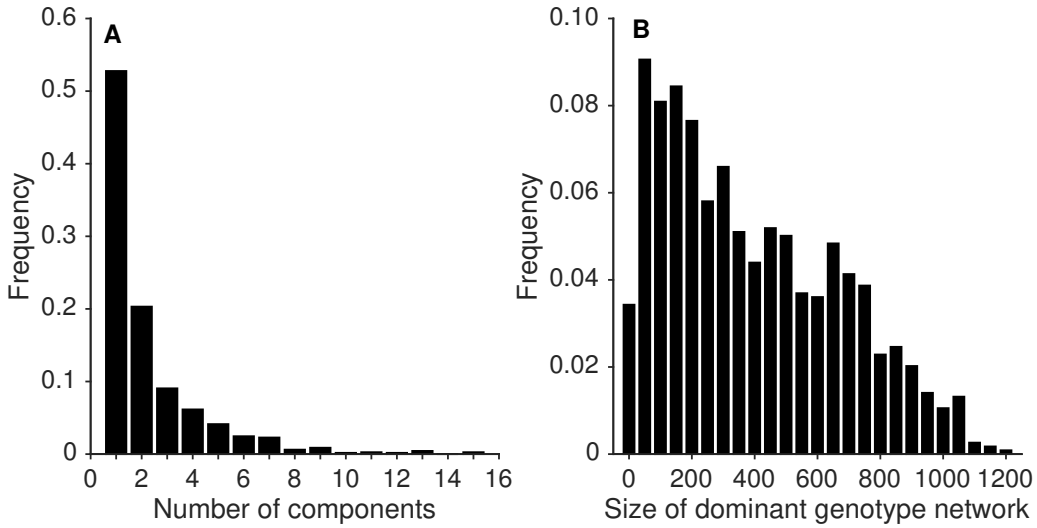
oligonucleotides via repeated rounds of mutation and selection. Biophysical models draw similar conclusions, but since they were not based on exhaustive measurements of binding affinity, they necessarily made several assumptions about landscape structure. These include the assumption that single nucleotide changes in a binding site result in small changes in binding affinity [328, 363, 364], that a binding site's constituent bases contribute additively to affinity [319, 359], and that binding affinity is a linear function of a site's mutational distance from the highest-affinity site [328, 363]. We and others have provided broad support for the first assumption, because the binding affinity of a sequence is strongly correlated with that of its mutational neighbors [25, 156]. In contrast, the latter assumptions are occasionally violated: Magnitude epistasis is common and landscapes sometimes have multiple peaks. These findings urge caution when incorporating additivity assumptions into models of TF-DNA interactions.

2.6.1 Caveats

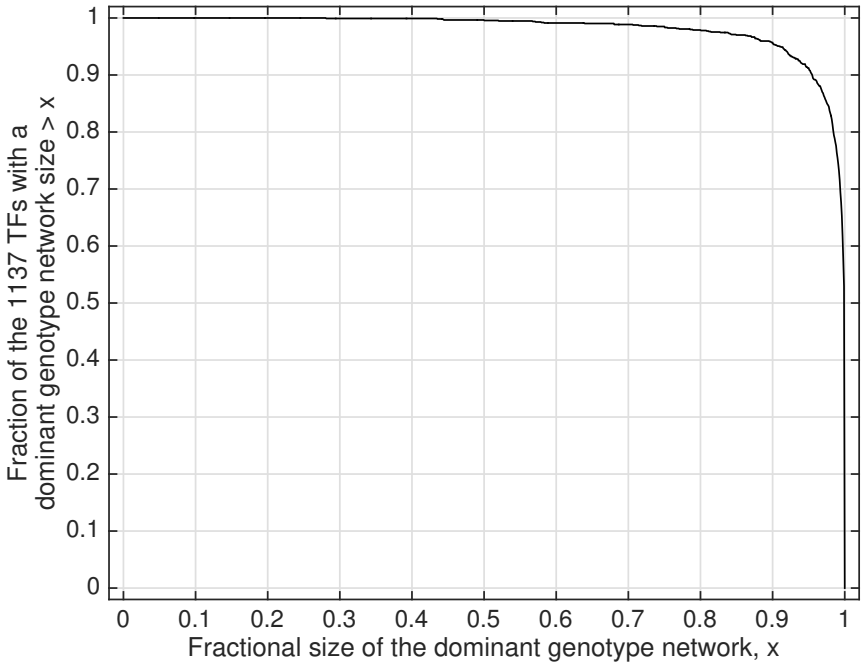
This study has some caveats that are worth highlighting. First, the landscapes we have constructed make several simplifying assumptions about transcriptional regulation. For instance, since they are based on an *in vitro* assay of TF binding, they do not capture many of the complexities of binding *in vivo*, such as epigenetic marks [25], chromatin context [365], local sequence context [366], or interactions with protein partners [334, 367], all of which are likely to affect landscape topography. Nevertheless, we observe significant correlations among our measures of landscape navigability and the *in vivo* abundance of high-affinity sites and gene expression levels, suggesting that these *in vitro* affinity landscapes provide meaningful information about TF binding *in vivo*. This is further supported by the observation that the genetic diversity of binding sites across yeast strains reflects the size of their landscape's global peaks. The second caveat is that we cannot completely rule out the possibility that some of the correlations we observe are spurious. While their consistency across different species, experiments, as well as diverse cell and tissue types is reassuring, only direct experimentation could conclusively determine whether the navigability of a binding affinity landscape affects the evolution of a TF's

binding sites *in vivo*. High-throughput laboratory evolution experiments with selection for transcription factor binding could shed light on which landscape parameterizations (τ and δ) best reflect binding *in vivo*, and could provide answers to long-standing questions about the influence of landscape navigability on adaptation, mutational robustness, and the predictability of evolution. Our current research is directed along these lines. Third, the metaphor of the adaptive landscape commonly evokes the fitness of an entire organism, and it is not immediately clear how TF binding affinity relates to organismic fitness. Even though examples—such as the evolution of an activator's binding site in the promoter of an antibiotic resistance gene [160]—have linked binding affinity to the fitness of an organism, such a link need not exist for all TFs. Moreover, for TF's regulating many genes, the relationship between binding affinity and fitness likely varies from binding site to binding site. However, we stress that this limitation is not restricted to binding affinity, but rather applies to the quantitative phenotypes measured to date in almost all empirical adaptive landscapes of proteins and RNAs, including catalytic activity, stability, or affinity to a target ligand [28, 56]. Fourth, while our measure of peak accessibility is inspired by the weak-mutation, strong-selection regime of population genetics, not all TF binding sites are under strong selection [338]. Our measure is therefore conservative, because relaxed selection for binding affinity would only make a landscape's peaks more accessible. Fifth, our landscape navigability measures assume that valleys cannot be crossed. This assumption may be violated if the mutation rate is high or if the population size is small. Valley crossing may also be facilitated by relaxed selection for TF binding, genetic hitchhiking, or by temporally changing environments [368]. A final caveat is that we only consider mutations in TF binding sites, even though mutations in the DNA binding domains of TFs are also important for regulatory evolution [357, 369, 370]. Data describing the effects of such mutations on a TF binding preferences are now available [357, 371, 372], but due to the hyper-astronomical size of protein space [41], they are far from exhaustive, capturing only a small subset of all possible mutations to a TF's DNA binding domain, and thus precluding a comprehensive analysis.

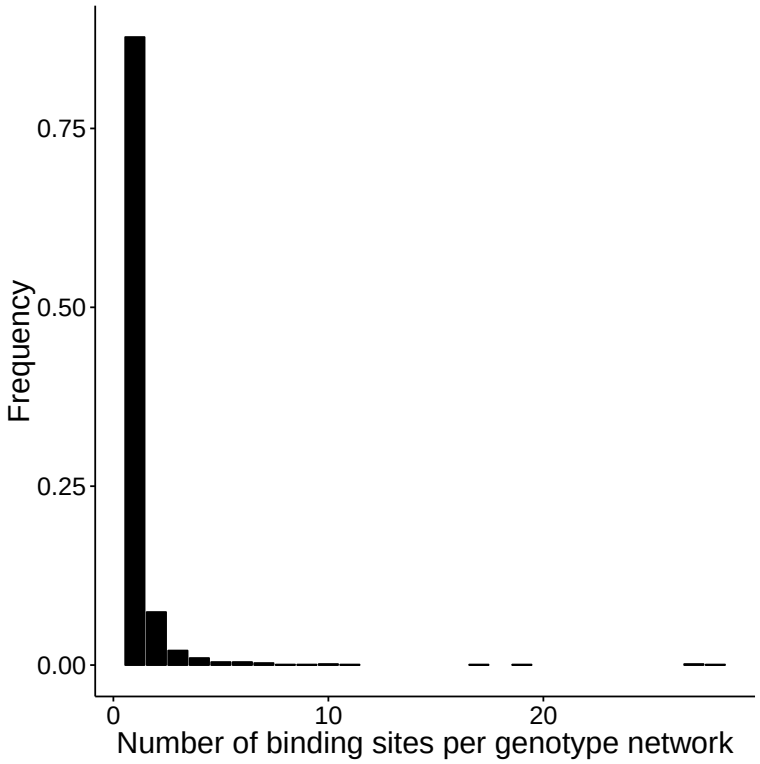
2.7 Supplementary figures



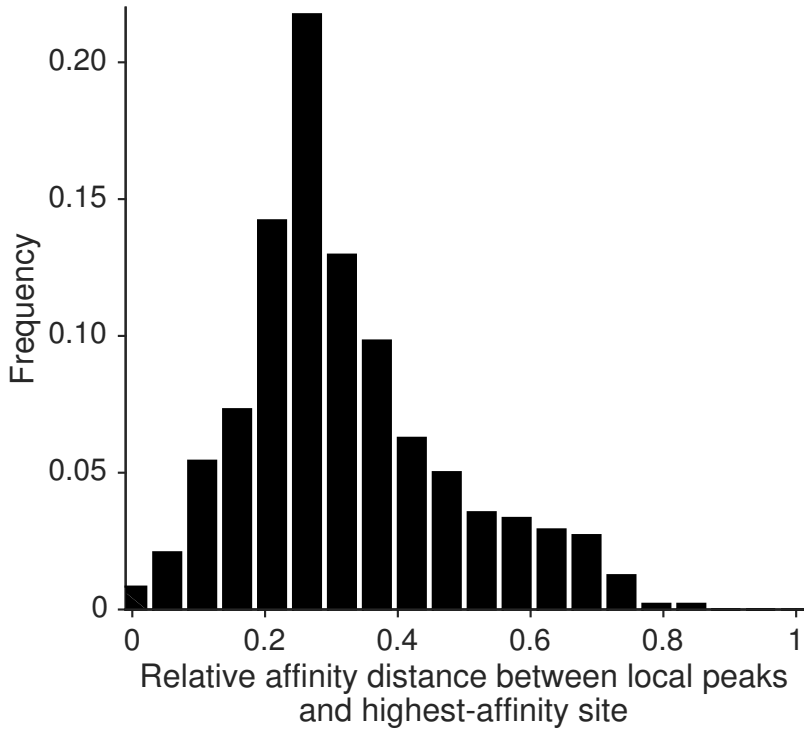
Supplementary Figure 2.1: Structural properties of the genotype networks of 1137 TFs. The distribution of (**A**) the number of components and (**B**) the size of the largest (i.e., dominant) connected component in the genotype networks of 1137 TFs.



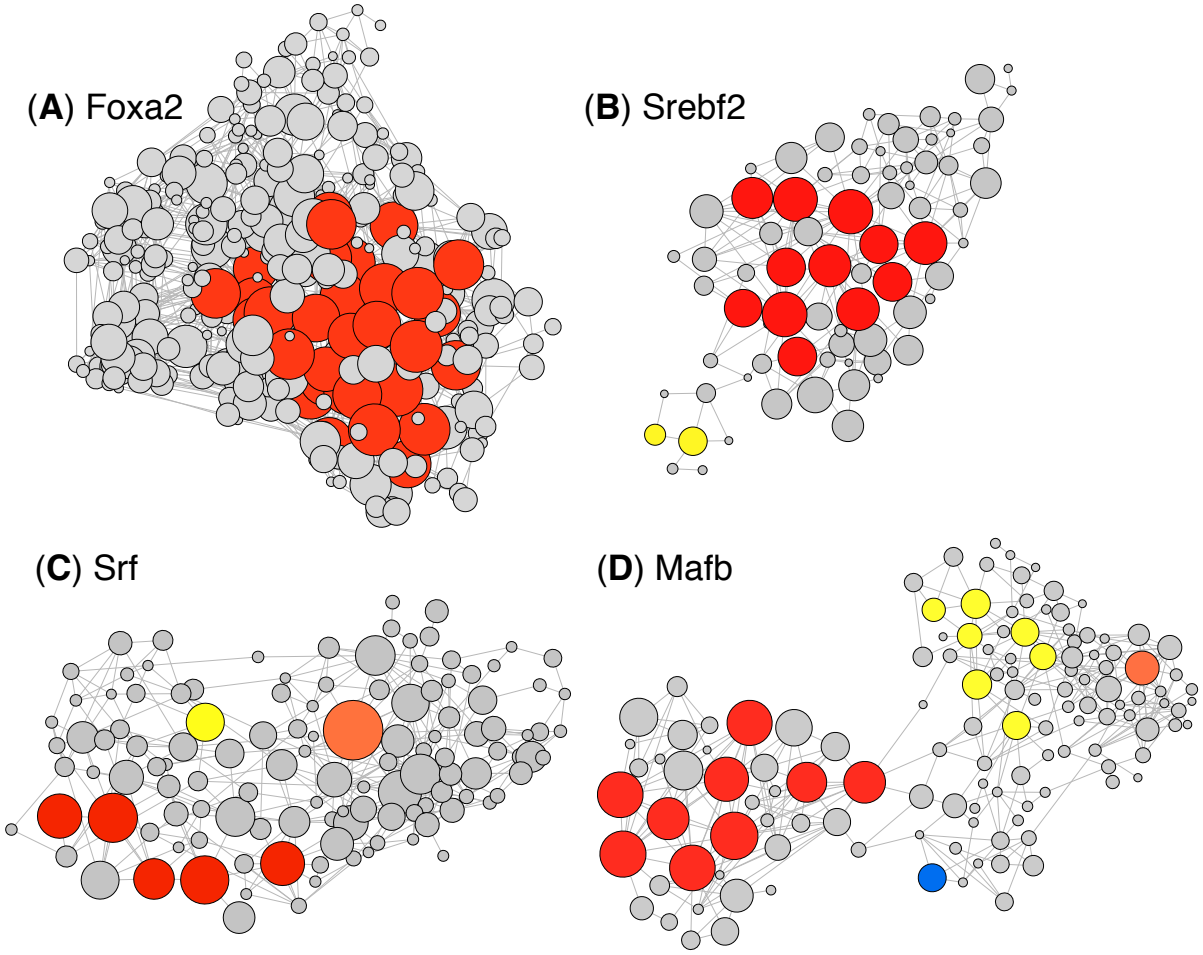
Supplementary Figure 2.2: Cumulative distribution of the fraction of TFs in our dataset that have a dominant genotype network comprising at least x% of the bound sequences. This figure shows that for the vast majority of TFs, the vast majority of bound sequences are in the dominant genotype network. For example, 96% of the 1137 dominant genotype networks comprise at least 90% of the sequences that bind the TF.



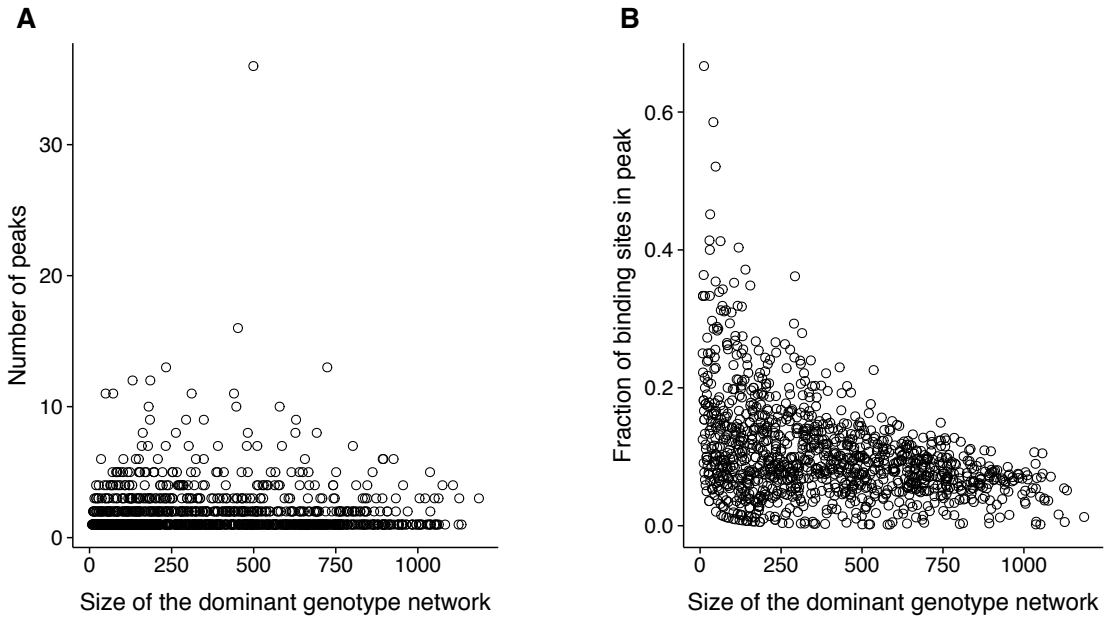
Supplementary Figure 2.3: The number of sequences per non-dominant genotype network is very small. Black bars show the distribution of the number of sequences contained in the non-dominant genotype networks of 1,137 TFs. 87.8% of these networks comprise a single sequence.



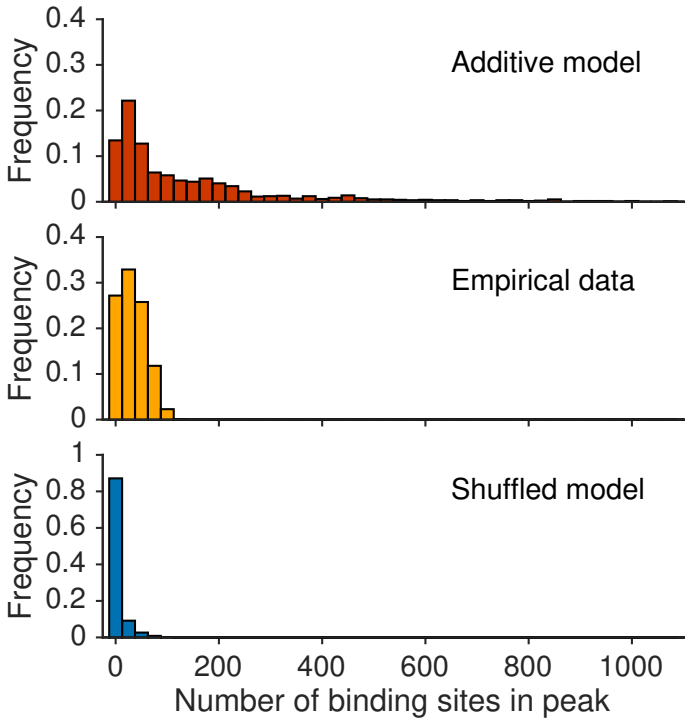
Supplementary Figure 2.4: Multiple peaks of unequal height. The distribution of the average normalized difference in binding affinity between the highest-affinity sequence in a landscape and the mean binding affinity of the landscape's local peaks, for 478 multi-peaked landscapes. This distance D_i of a local peak i is calculated as $E_g - \frac{1}{N_1}\sum_j E_{ij}$, where E_g is the binding affinity of the highest-affinity site in the global peak, N_i is the number of sites in local peak i, and E_{ij} is the binding affinity of site j in local peak i. All values are normalized by 0.15, which is the maximum range of affinity when the binding threshold is $\tau = 0.35$. Non-normalized values range from 1.3×10^{-3} to 0.125.



Supplementary Figure 2.5: Visualization of adaptive landscapes that have different numbers of peaks. Representative examples of landscapes with (**A**) one, (**B**) two, (**C**) three, and (**D**) four peaks. Vertices represent bound sequences, vertex sizes represent binding affinity, and vertex colours represents peak membership. Red indicates a global peak and grey indicates bound sequences that do not belong to any peak. Notice that peaks are usually organized into broad plateaus comprising binding sites with similar binding affinity (Supplementary Fig. 2.7), that is with values in affinity above a threshold $E_g - \delta$, where E_g is the affinity of the highest-affinity site in the peak and δ is the noise threshold ('Methods').

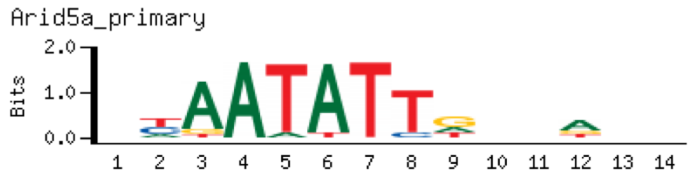


Supplementary Figure 2.6: Small landscapes are not necessarily more navigable than large landscapes. The relationship between the size of the dominant genotype network and (**A**) the number of peaks (Spearman's rank correlation coefficient = -0.07, P value $< 1.9 \times 10^{-2}$), and (**B**) the relative size of the global peak for 1,137 TFs (Spearman's rank correlation coefficient = -0.34, P value $< 2.2 \times 10^{-16}$).



Supplementary Figure 2.7: Global peaks are usually organized into broad plateaus. The distribution of the number of binding sites per global peak for the 1,137 empirical adaptive landscapes (yellow) and the additive (red) and rugged (blue) null models.

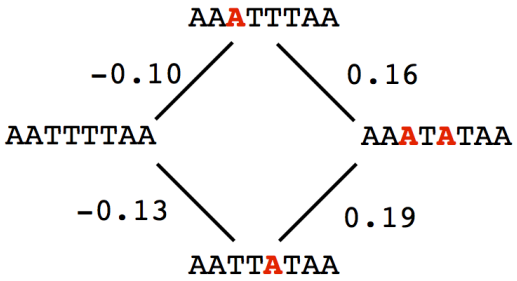
(A)



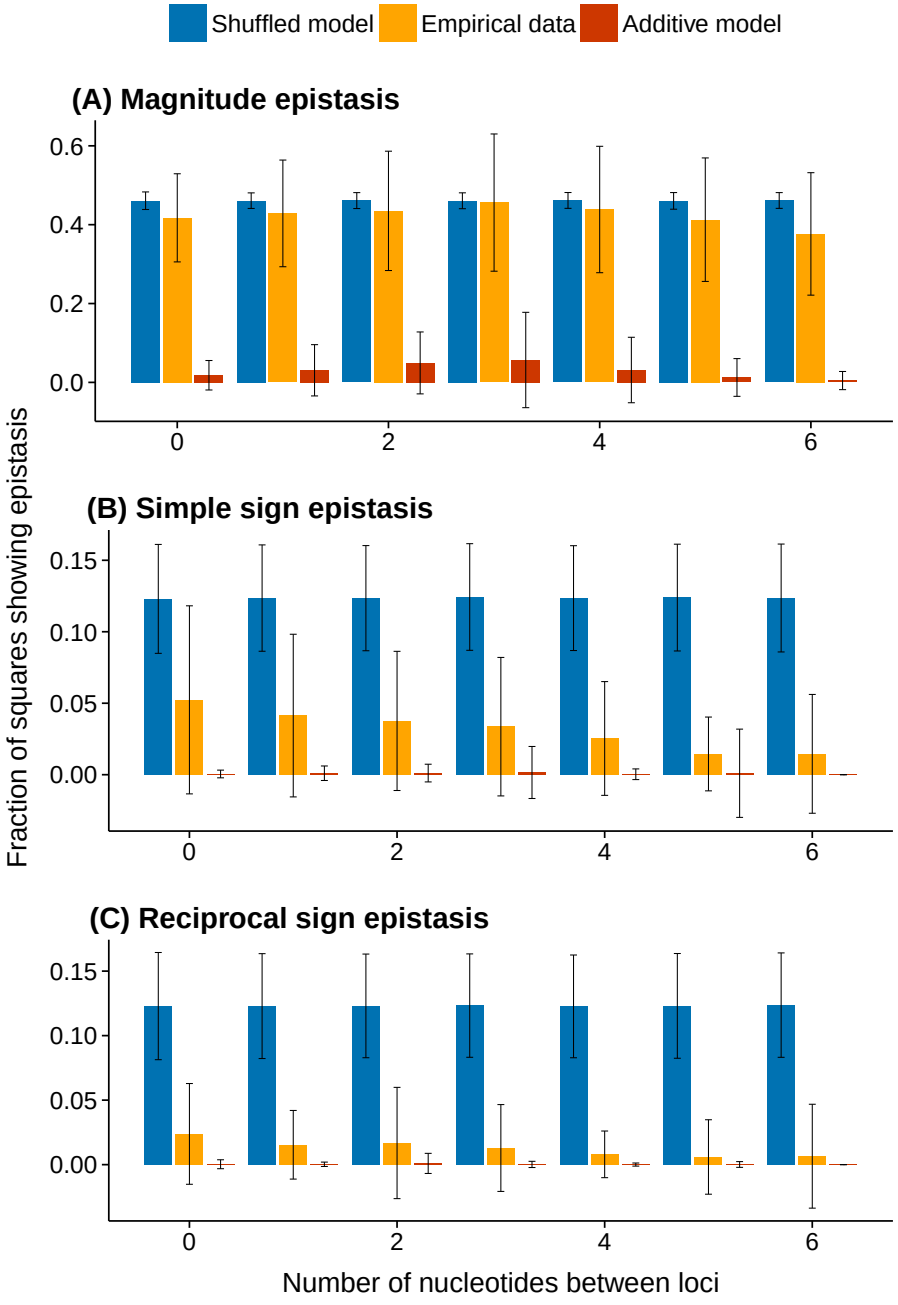
(B)

A:	0.177	0.160	0.852	0.961	0.056	0.930	0.019	0.041	0.233	0.339	0.285	0.572	0.285	0.342
C:	0.428	0.324	0.026	0.003	0.004	0.005	0.014	0.087	0.027	0.352	0.134	0.078	0.184	0.230
G:	0.229	0.032	0.073	0.010	0.007	0.007	0.006	0.039	0.519	0.184	0.268	0.190	0.262	0.153
T:	0.165	0.482	0.047	0.024	0.931	0.057	0.959	0.832	0.218	0.123	0.311	0.159	0.268	0.272
I(i):	0.118	0.384	**1.183**	**1.724**	**1.589**	**1.579**	**1.703**	**1.101**	0.399	0.120	0.062	0.375	0.019	0.058
Pos:	1	2	**3**	**4**	**5**	**6**	**7**	**8**	9	10	11	12	13	14

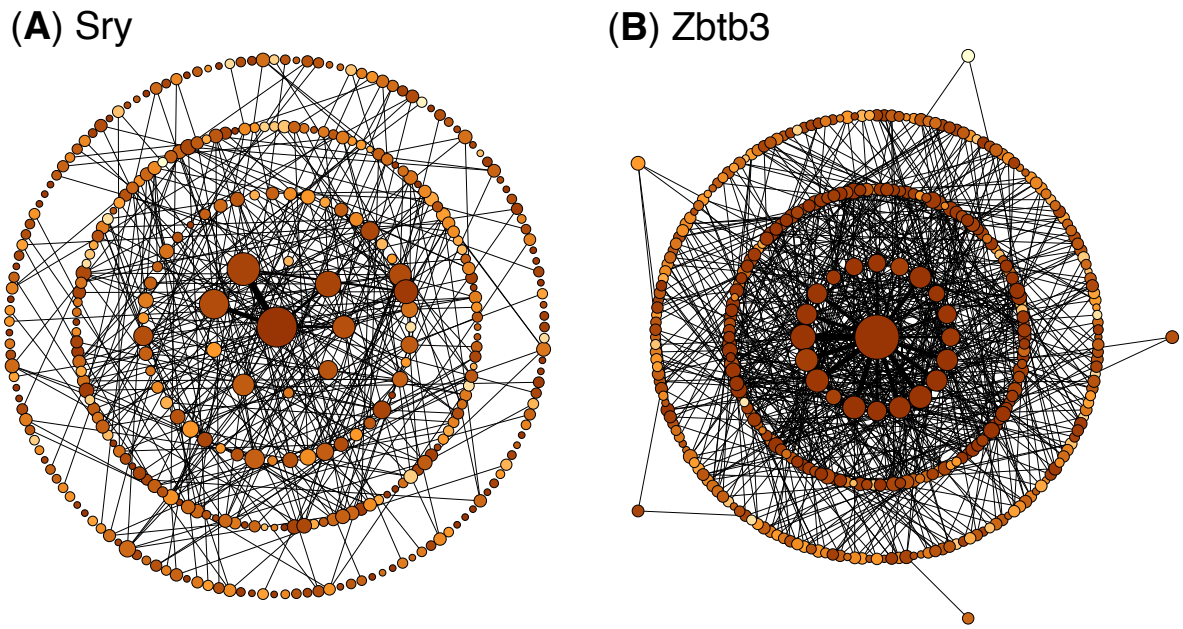
(C)



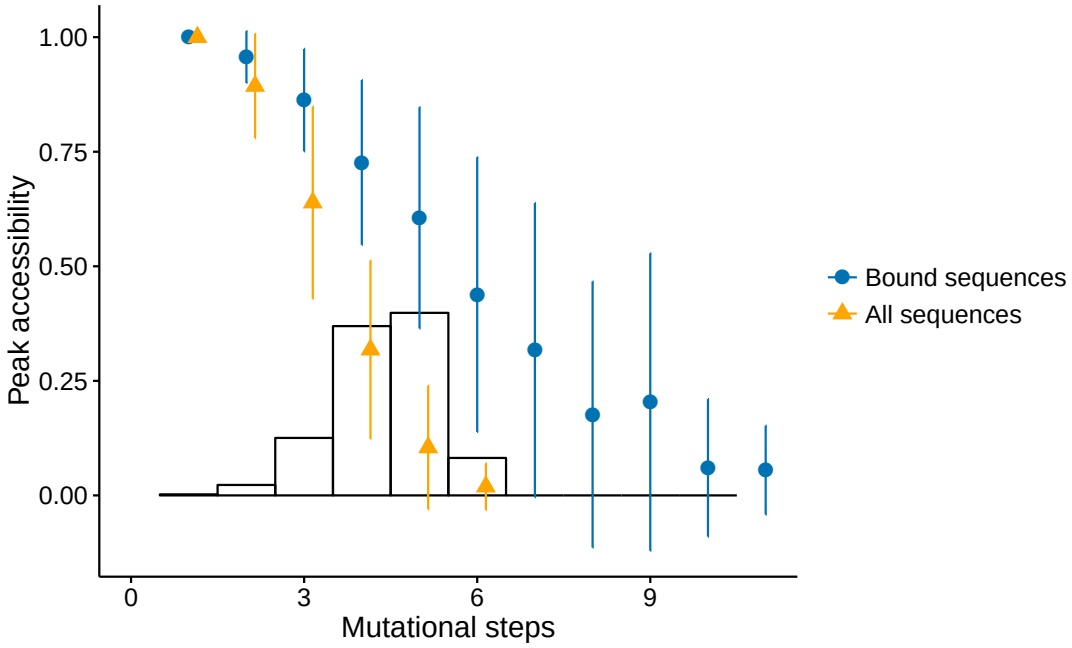
Supplementary Figure 2.8: Why epistasis occasionally appears in the additive null model. (**A**) Sequence logo of (**B**) the position weight matrix (PWM) for the mouse TF Arid5a downloaded from UniPROBE [322]. The information content (Eq. (2.8), 'Methods') per position is also shown. Positions in bold face indicate the largest contiguous subset of positions with information content above 0.5, which we use to calculate the effective width of a PWM (Supplementary section 2.5.6.4). (**C**) A square that exhibits reciprocal sign epistasis due to the sliding-window approach for scoring sequences in the additive null model. Edge labels indicate the change in binding affinity between a pair of sequences. For the sequence AATTTTAA, the highest-scoring match occurs when it is aligned with positions 3 through 10 in the PWM, whereas for the other three sequences, the highest-scoring match occurs when they are aligned with positions 2 through 9 in the PWM. In this example, the noise threshold δ is 0.09 and the magnitude of epistasis (Eq. (2.3), 'Methods') is 0.29.



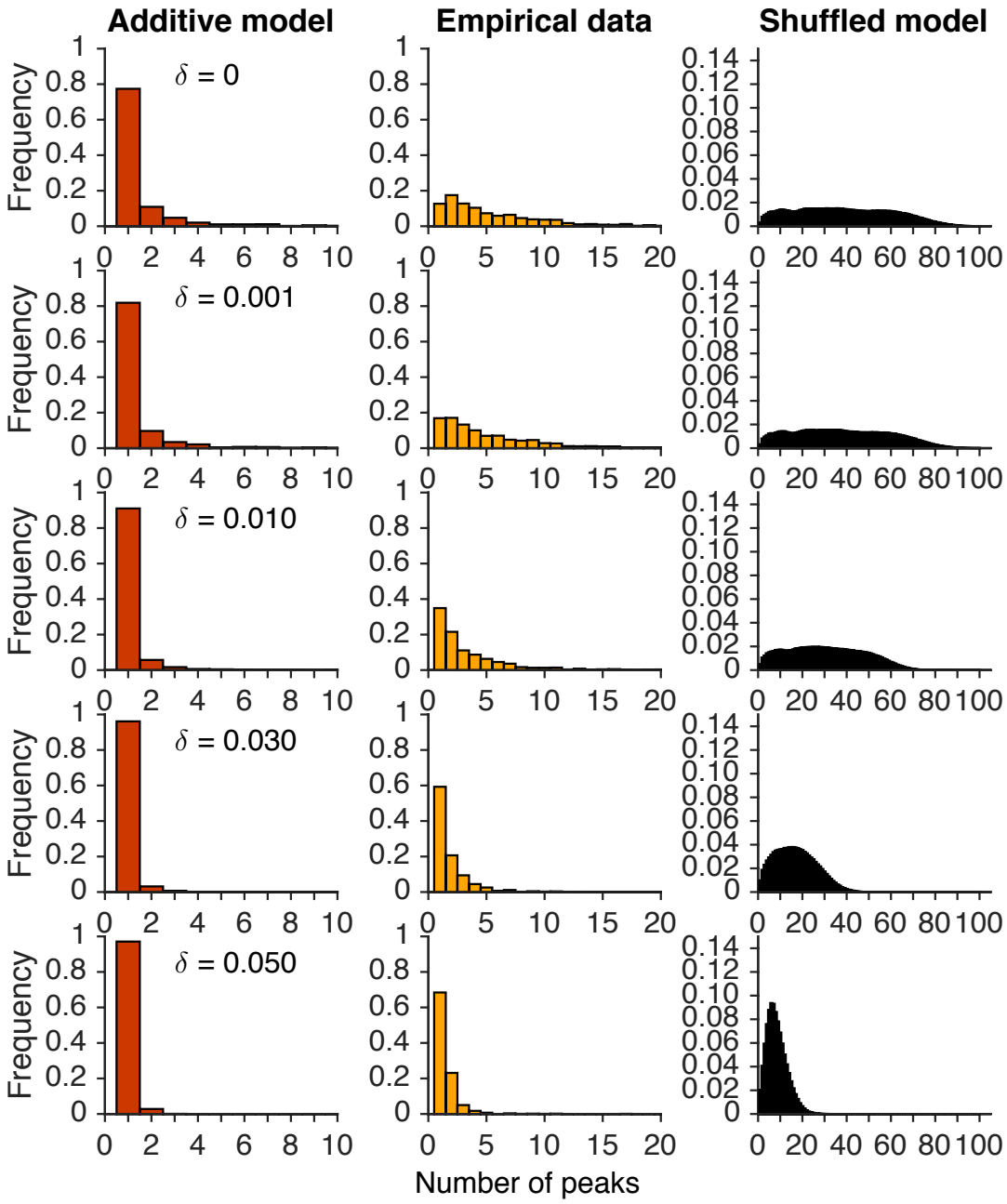
Supplementary Figure 2.9: Sign epistasis preferentially occurs among nucleotides that are near one another in the binding site. Bar plots of the levels of (**A**) magnitude (**B**) simple sign, and (**C**) reciprocal sign epistasis as a function of the distance between the two mutations in the TF binding site. The height of the bars represents the median of the data, while the error bars represent the standard deviation. For each TF, the level of epistasis plotted for the random landscapes is an average of 1,000 shuffled landscapes. To facilitate a direct comparison with Jolma *et al.* [324], squares that include indels are excluded from this analysis.



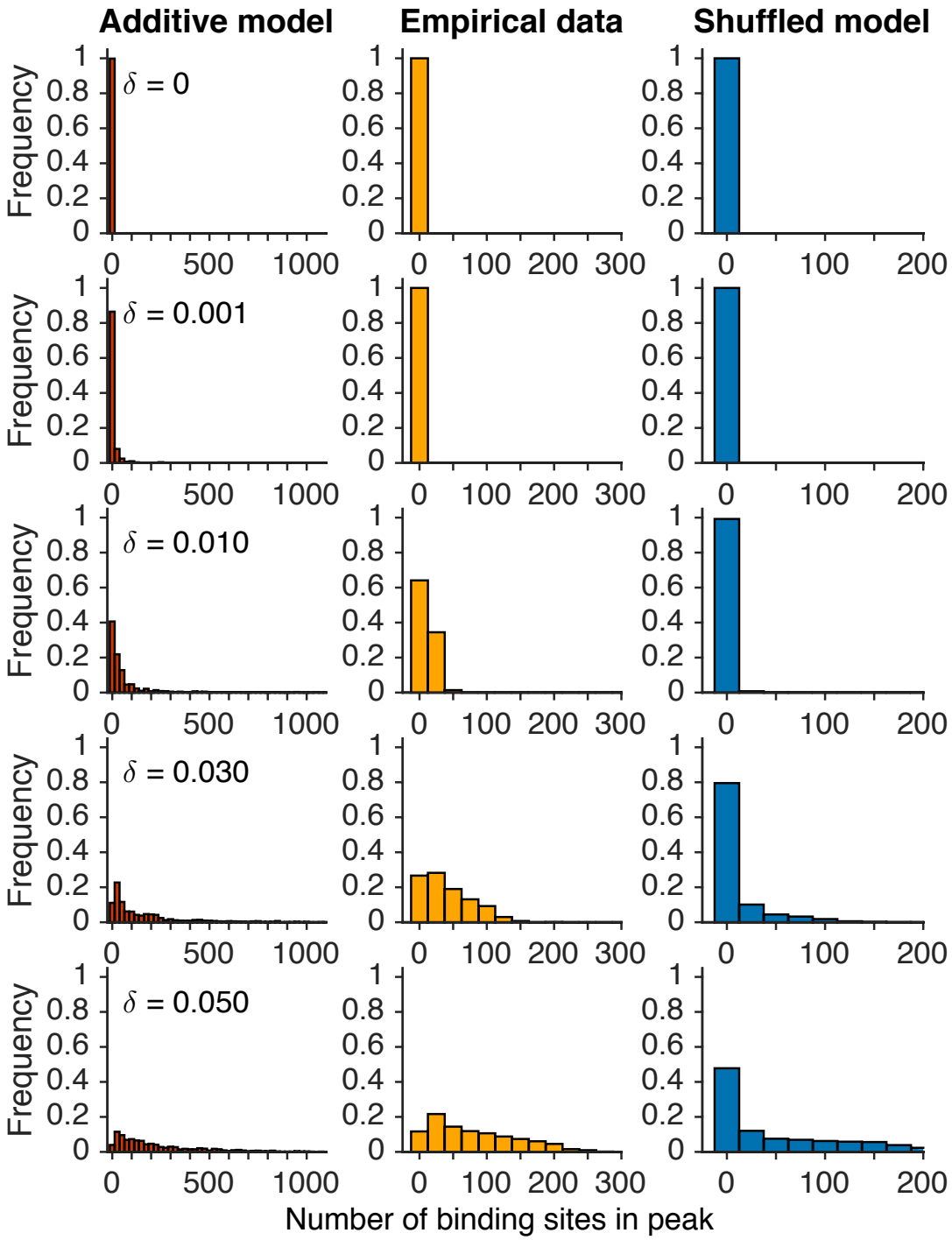
Supplementary Figure 2.10: Visualization of two global peaks that vary in their accessibility. The murine TFs (**A**) Sry and (**B**) Zbtb3 bind a similar number of sequences (356 and 352, respectively), but these sequences differ dramatically in their ability to evolve higher binding affinity through small mutations. Each visualization depicts the genotype networks of all bound sequences that are within four mutations of the highest-affinity site (centre vertex), i.e., the summit of the landscape, in a layout where each concentric ring corresponds to a distinct mutational distance from the summit. Edge width and vertex size indicate the number of accessible mutational paths (i.e., paths that access the summit via monotonic increases in binding affinity) that include the edge or vertex. Vertex color indicates binding affinity, as in Fig. 2.1a.



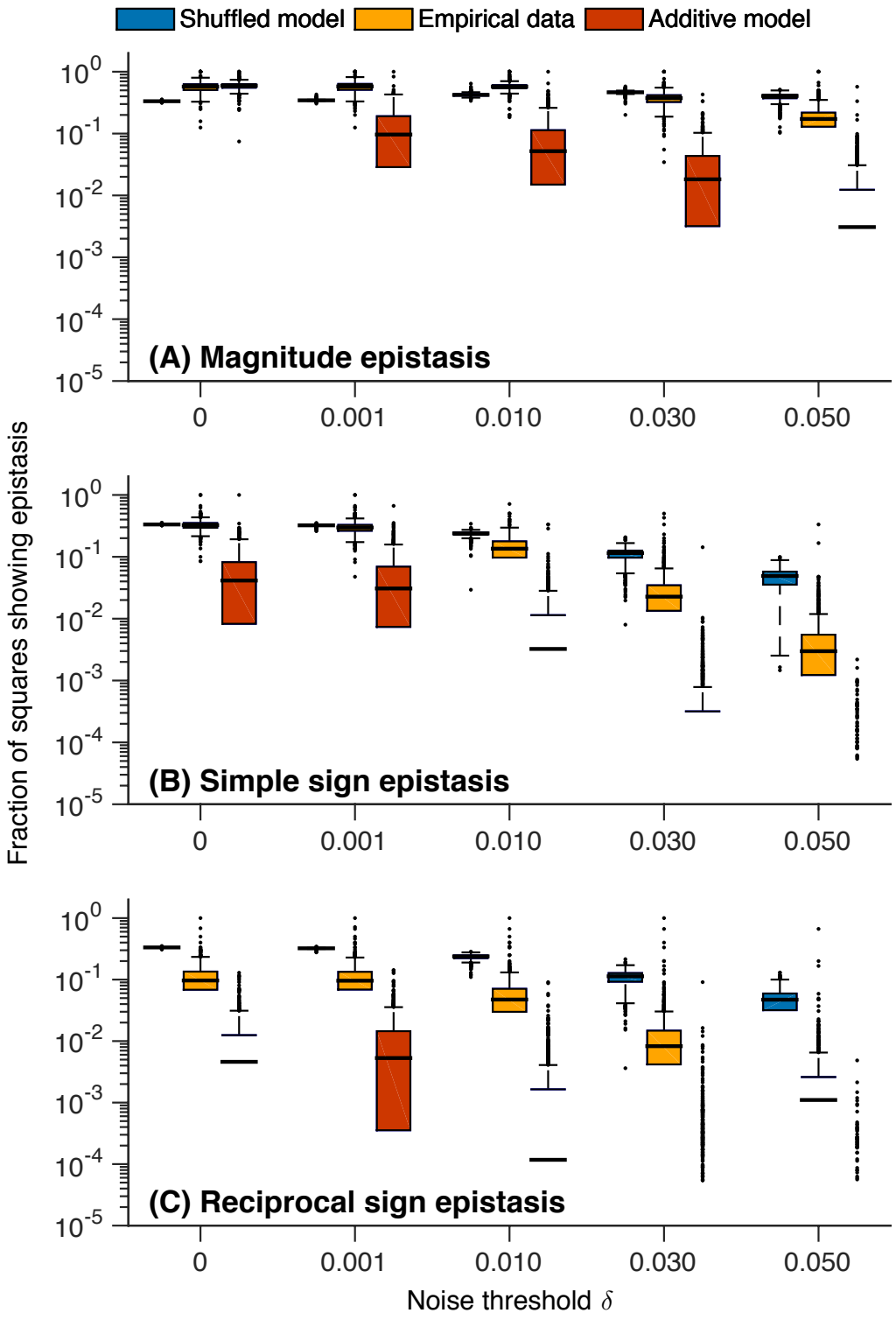
Supplementary Figure 2.11: Peak accessibility is reduced by the inclusion of unbound sequences. Data correspond to all 1,137 TFs. Symbols show the mean, and error bars show the standard deviation. The histogram shows the distribution of the lengths of the shortest mutational paths from bound sequences (i.e., $\tau > 0.35$) to the global peaks of the 1,137 TFs' landscapes, when unbound sequences (i.e., $\tau \leq 0.35$) are considered. Note that there is no shortest mutational path longer than 6 mutational steps in this case, in contrast to when unbound sequences are not considered (cf. Figure 2.2c).



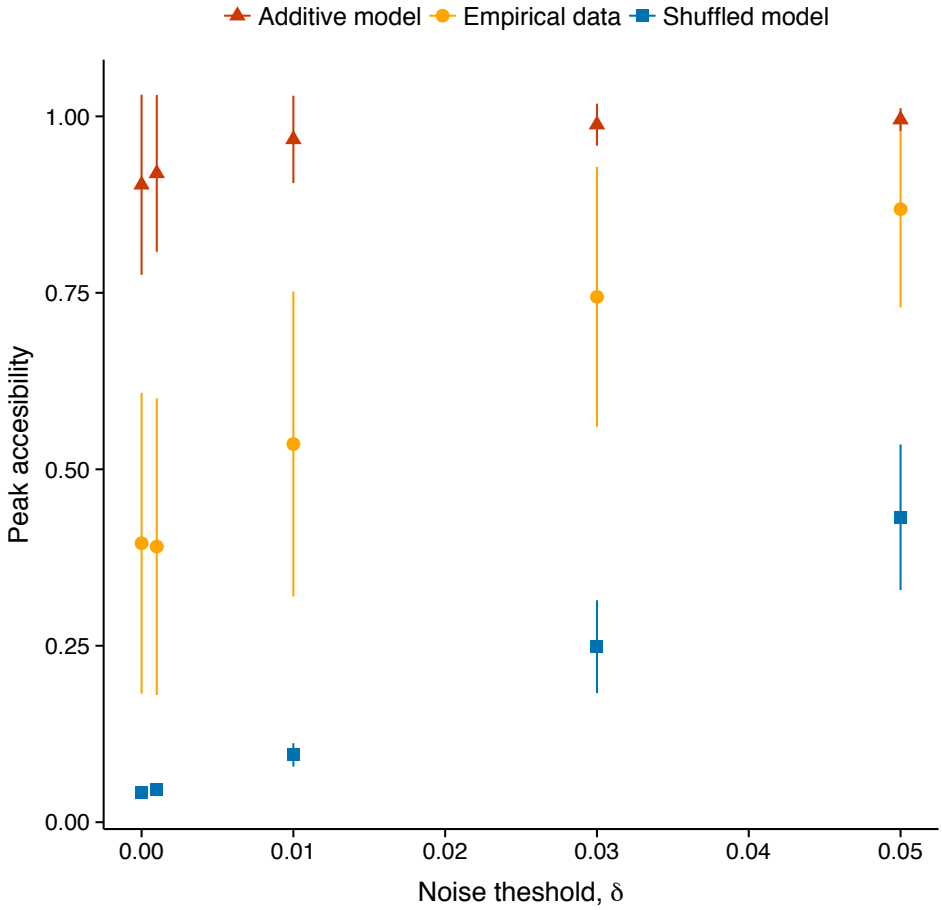
Supplementary Figure 2.12: The number of peaks per landscape decreases as the noise threshold increases. These data represent a sensitivity analysis of the trends presented in Fig. 2.2a. Each panel shows the distribution of the number of peaks per landscape for the additive model (left column), empirical data (middle column), and shuffled model (right column). Each row of panels corresponds to a different noise threshold δ.



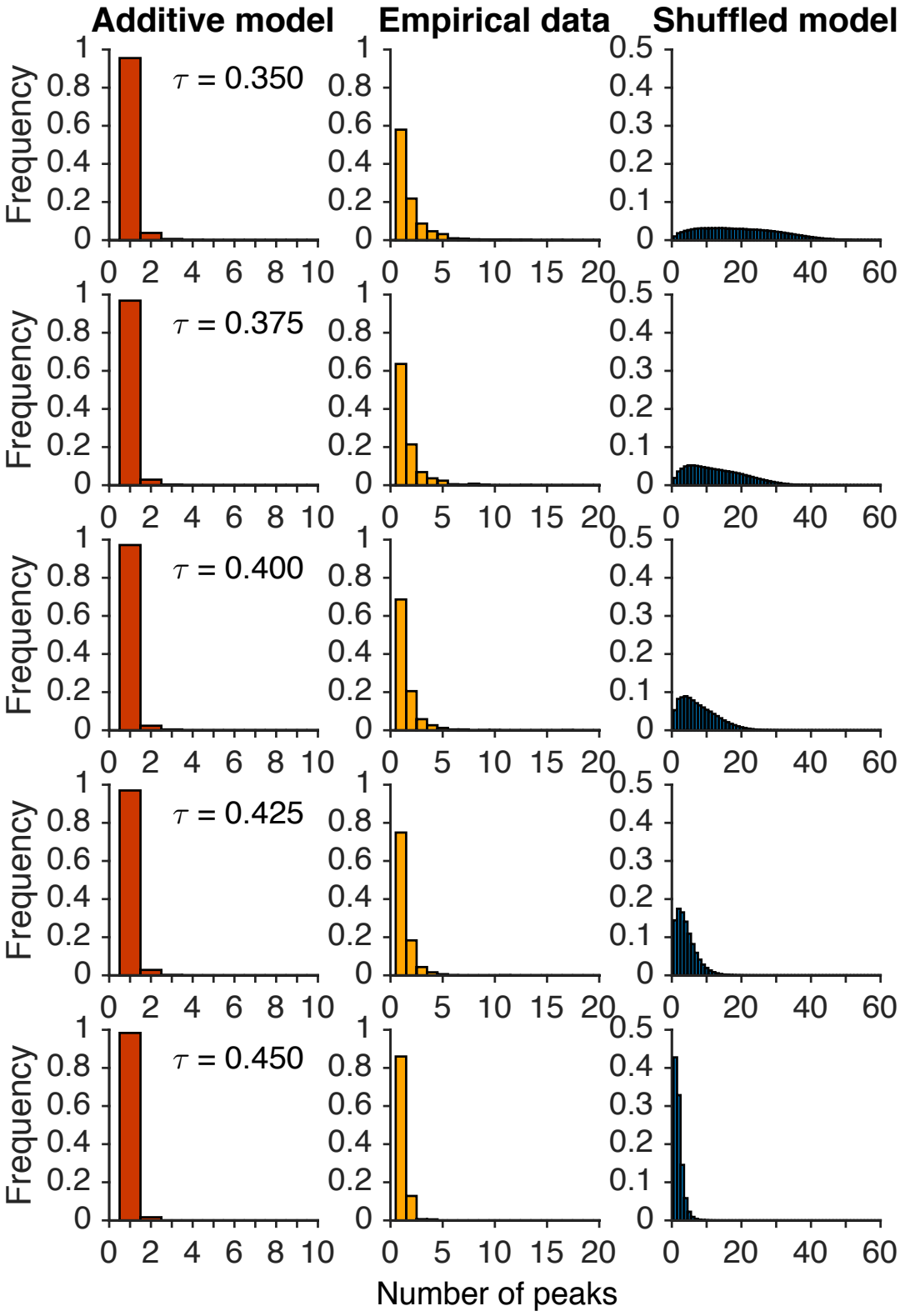
Supplementary Figure 2.13: Global peak breadth increases as the noise threshold increases. These data represent a sensitivity analysis of the trends presented in Supplementary Fig. 2.7. Each panel shows the distribution of the number of binding sites per global peak (i.e., global peak breadth) for the additive model (left column), empirical data (middle column), and shuffled model (right column). Each row of panels corresponds to a different noise threshold δ.



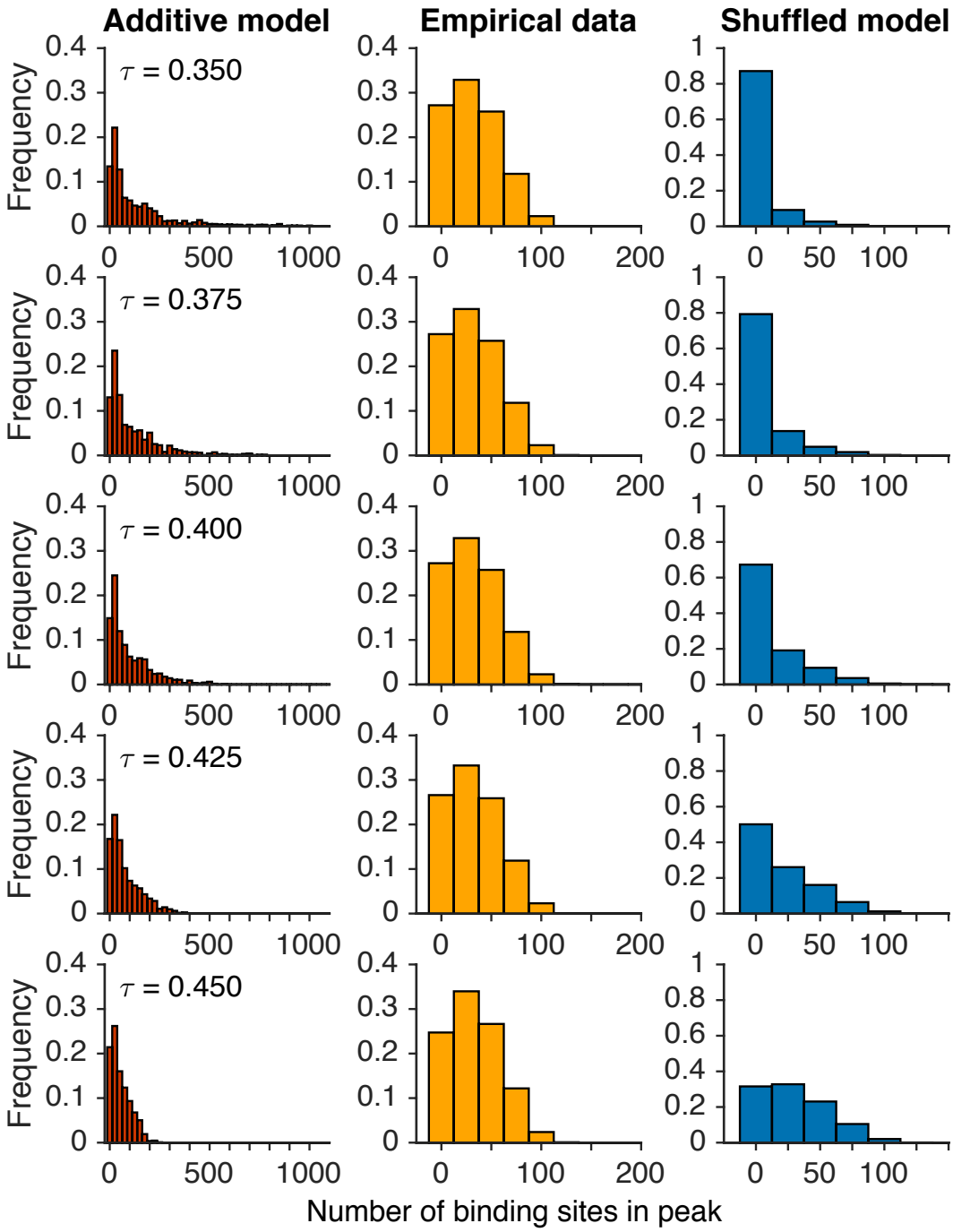
Supplementary Figure 2.14: Sign epistasis in the empirical data is intermediate to that of the additive and shuffled models for all noise thresholds. These data represent a sensitivity analysis of the trends presented in Fig. 2.2b. Boxplots of (**A**) magnitude epistasis, (**B**) simple sign epistasis, and (**C**) reciprocal sign epistasis as a function of the binding affinity threshold τ. The thick horizontal line in the middle of each box represents the median of the data, while the bottom and top of each box represent the 25th and 75th percentiles, respectively. Note the logarithmic scale of the y axes, which obscures data points with zero epistasis. The absence of the thick horizontal line indicates that the median of the distribution is below the lowest value of the y axis.



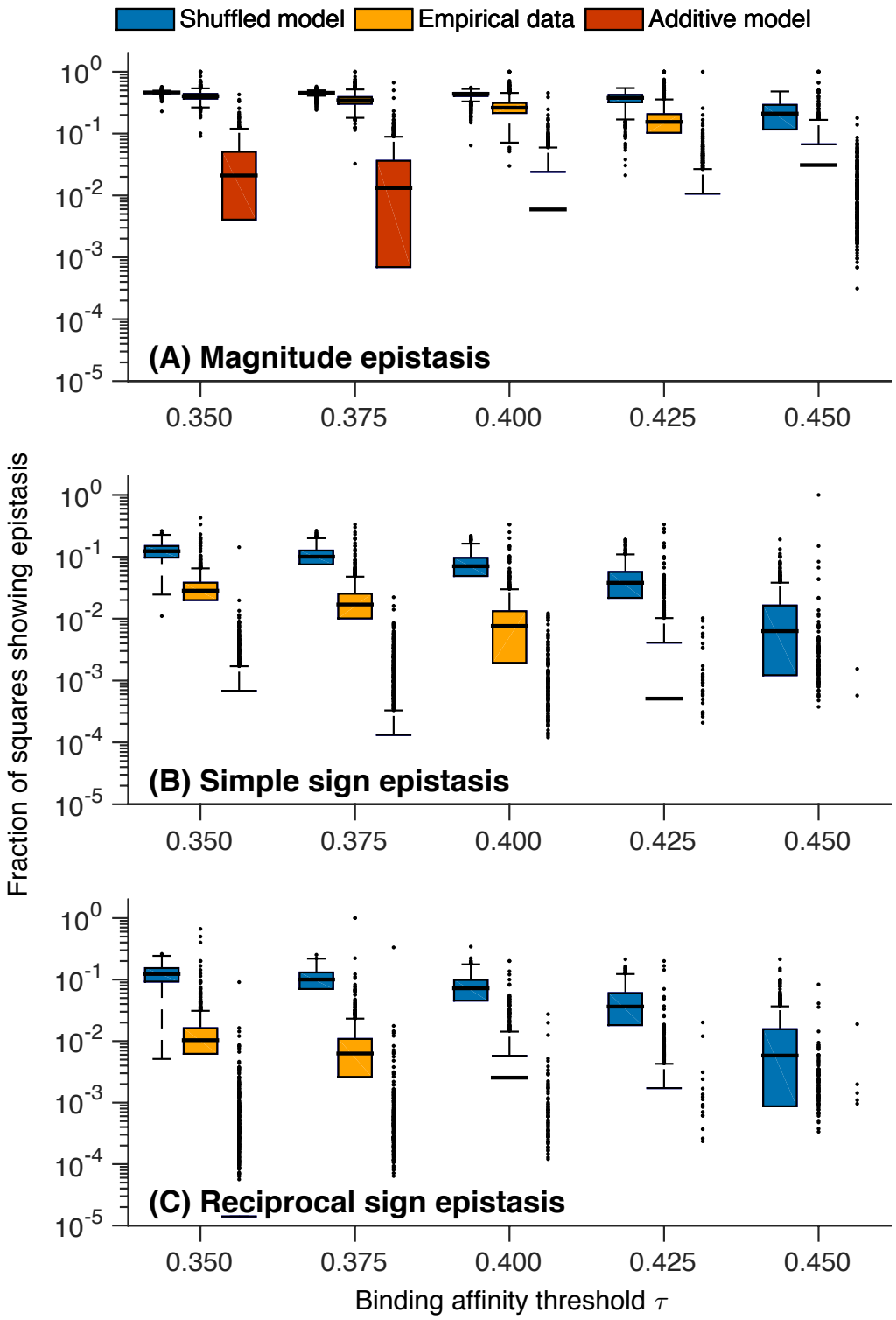
Supplementary Figure 2.15: Accessibility of the highest affinity site increases as the noise threshold increases. These data represent a sensitivity analysis of the trends presented in Fig. 2.2c for mutational paths of length 4, which are the most abundant in our dataset. Data points depict the mean peak accessibility in the empirical data and the two null models (see legend) as a function of the noise threshold δ. Error bars depict one standard deviation.



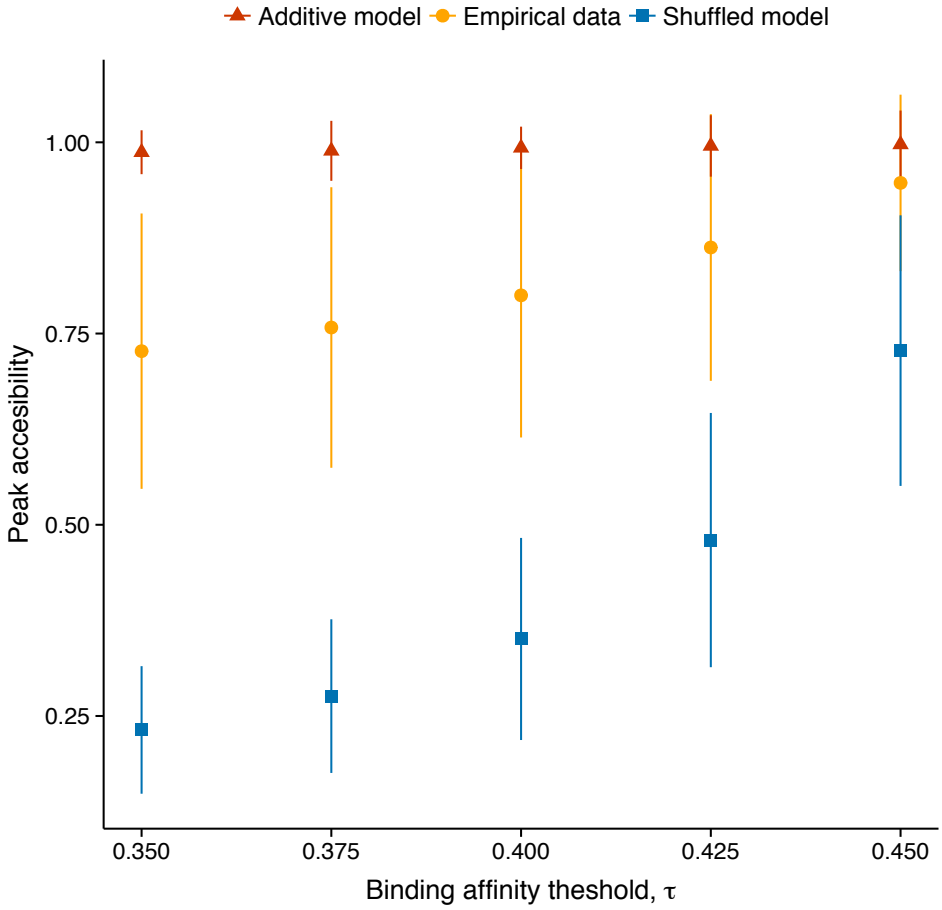
Supplementary Figure 2.16: The number of peaks per landscape decreases as the binding affinity threshold increases. These data represent a sensitivity analysis of the trends presented in Fig. 2.2a. Each panel shows the distribution of the number of peaks per landscape for the additive model (left column), empirical data (middle column), and shuffled model (right column). Each row of panels corresponds to a different binding affinity threshold τ.



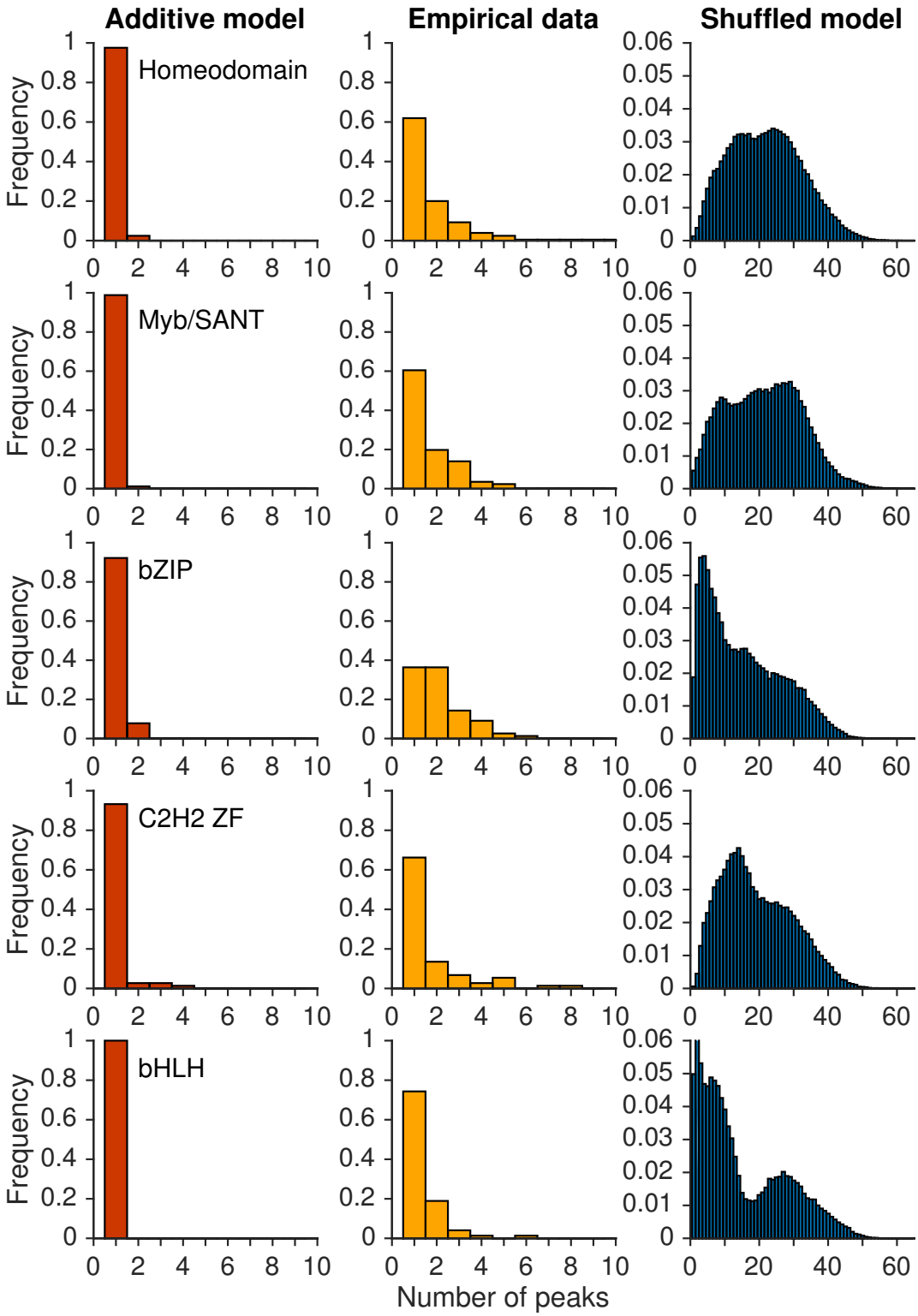
Supplementary Figure 2.17: The empirical distribution of global peak breadth does not vary with the binding affinity threshold. These data represent a sensitivity analysis of the trends presented in Supplementary Fig. 2.7. Each panel shows the distribution of the number of binding sites per global peak (i.e., global peak breadth) for the additive model (left column), empirical data (middle column), and shuffled model (right column). Each row of panels corresponds to a different binding affinity threshold τ.



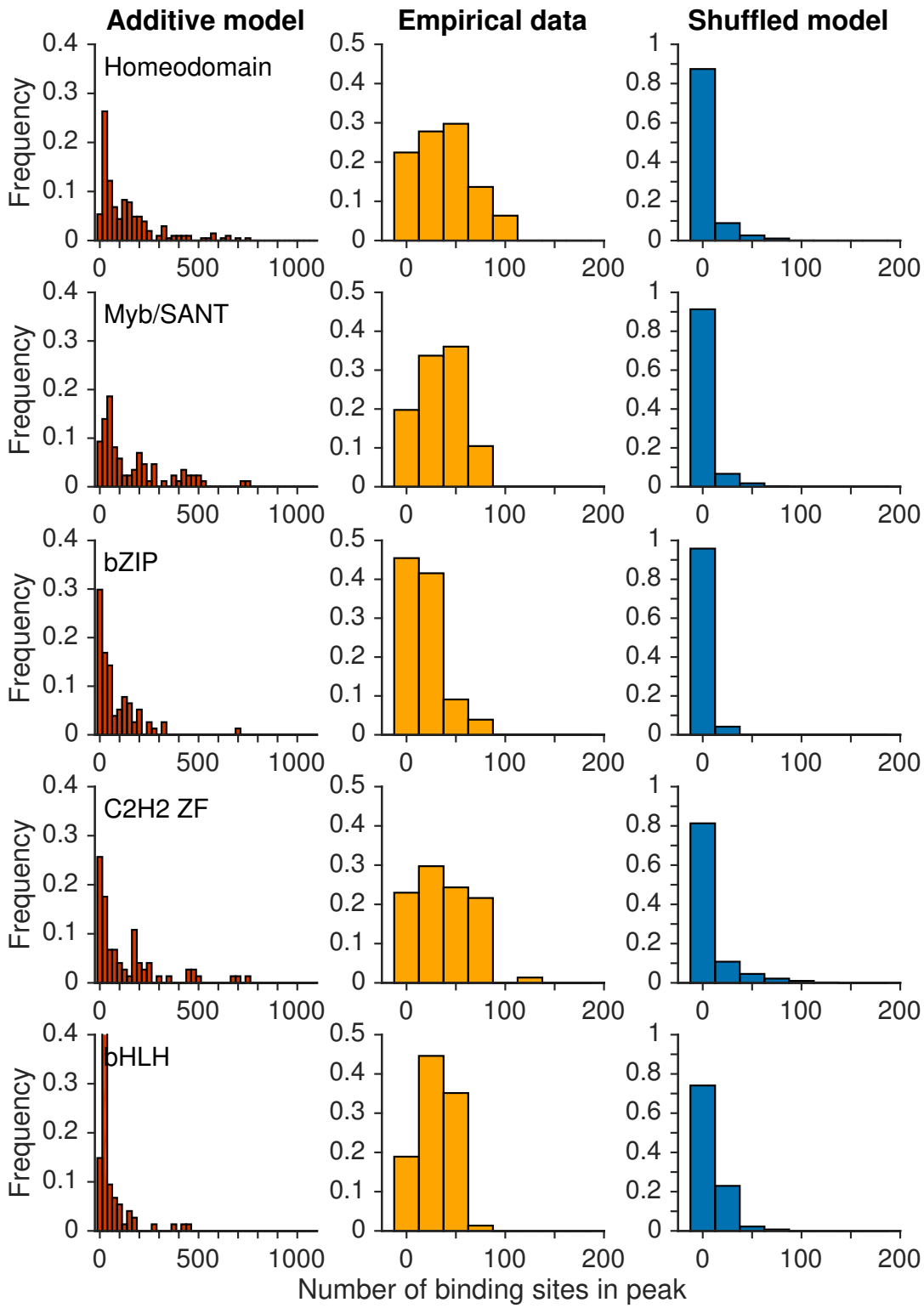
Supplementary Figure 2.18: Epistasis in the empirical data is intermediate to that of the additive and shuffled models for all affinity thresholds. These data represent a sensitivity analysis of the trends presented in Fig. 2.2b. Boxplots of (**A**) magnitude epistasis, (**B**) simple sign epistasis, and (**C**) reciprocal sign epistasis as a function of the binding affinity threshold τ. The thick line in the middle of each box represents the median of the data, while the bottom and top of each box represent the 25th and 75th percentiles, respectively. Note the logarithmic scale of the y axes, which obscures data points with zero epistasis. The absence of the thick line indicates that the median of the distribution is below the lowest value of the y axis.



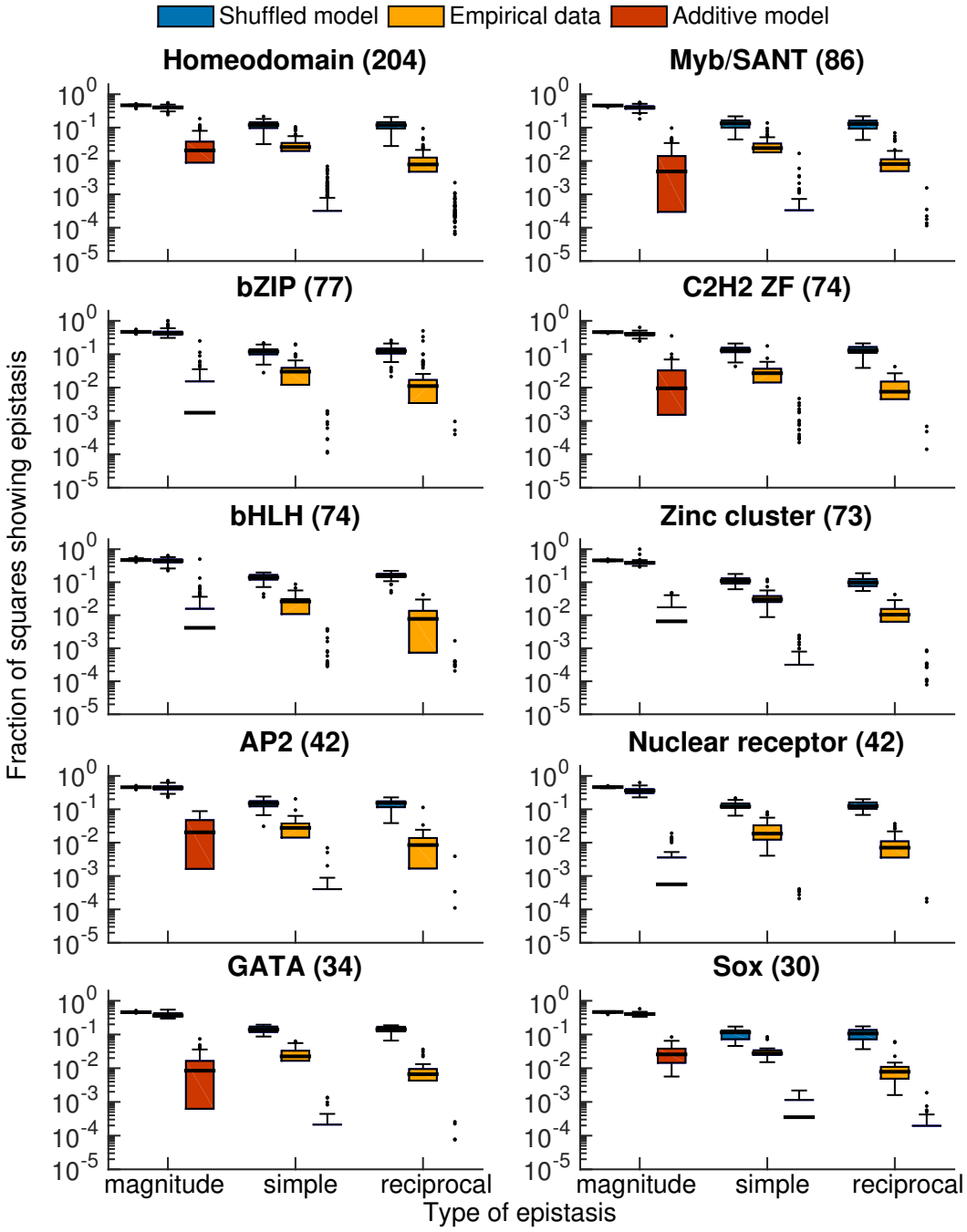
Supplementary Figure 2.19: Accessibility of the highest-affinity site increases as the binding affinity threshold increases. The data shown represent a sensitivity analysis of the trends presented in Fig. 2.2c for mutational paths of length 4, which are the most abundant in our dataset. Data points depict the mean peak accessibility in the empirical data and the two null models (see legend), as a function of the binding affinity threshold τ. Error bars depict one standard deviation.



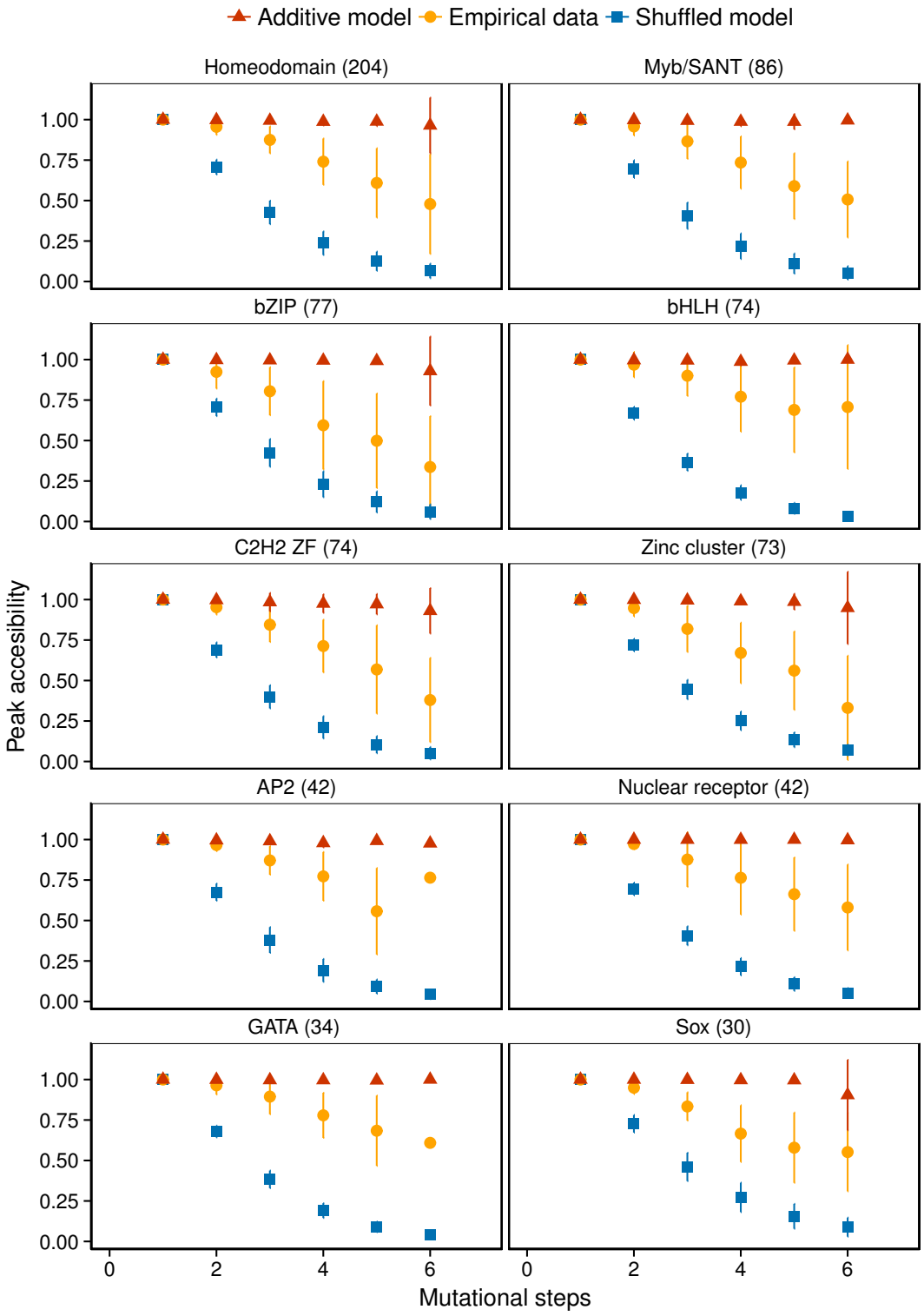
Supplementary Figure 2.20: The number of peaks is intermediate to that of the additive and shuffled models across DNA binding domains. These data represent a sensitivity analysis of the trends presented in Fig. 2.2a. Each panel shows the distribution of the number of peaks per landscape for the additive model (left column), empirical data (middle column), and shuffled model (right column). Each row of panels corresponds to a different DNA binding domain family, chosen because they are the five most prominent in our dataset.



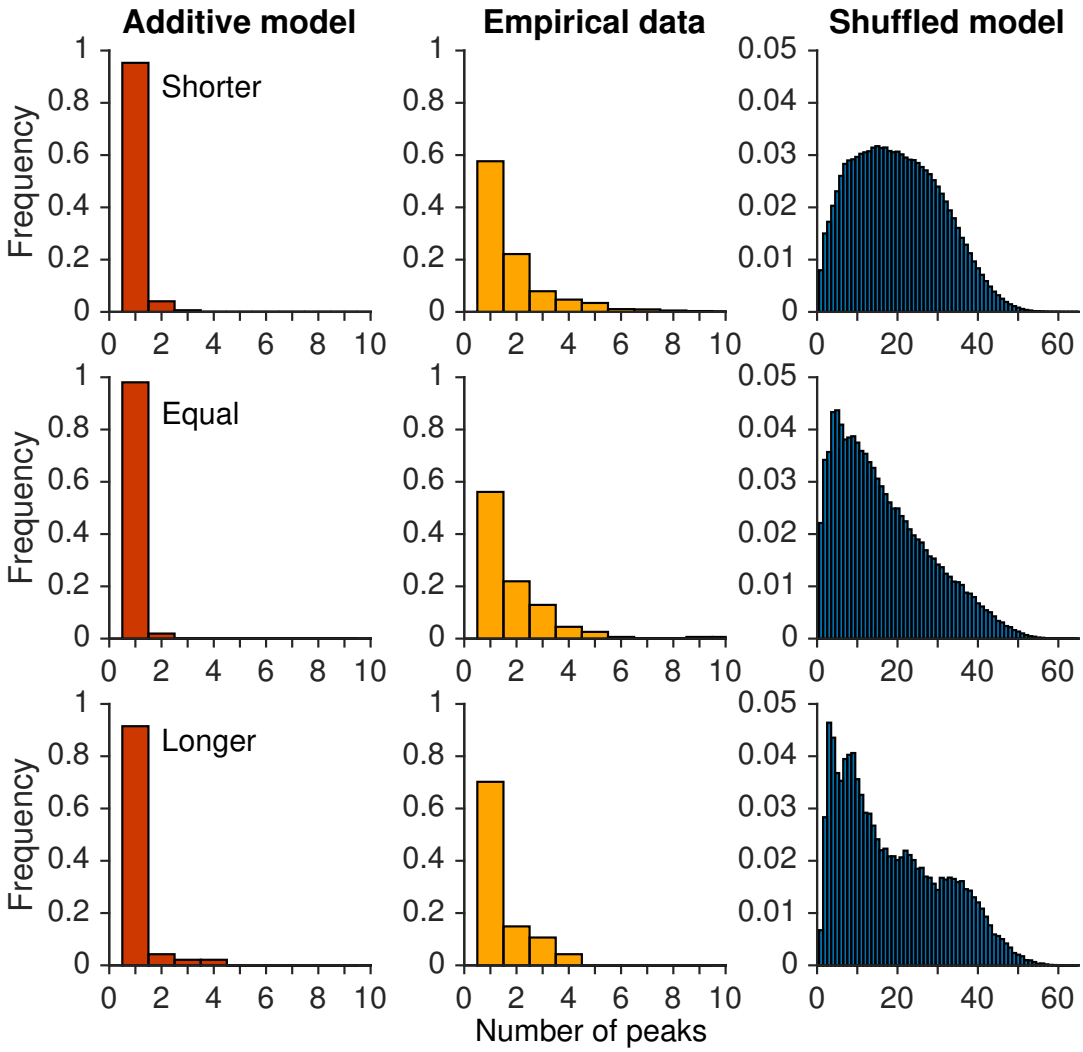
Supplementary Figure 2.21: The size of the global peak is intermediate to that of the additive and shuffled models across DNA binding domains. These data represent a sensitivity analysis of the trends presented in Supplementary Fig. 2.7. Each panel shows the distribution of the number of binding sites in the highest peak (i.e. global peak breadth) for the additive model (left column), empirical data (middle column), and shuffled model (right column). Each row of panels corresponds to a different DNA binding domain family.



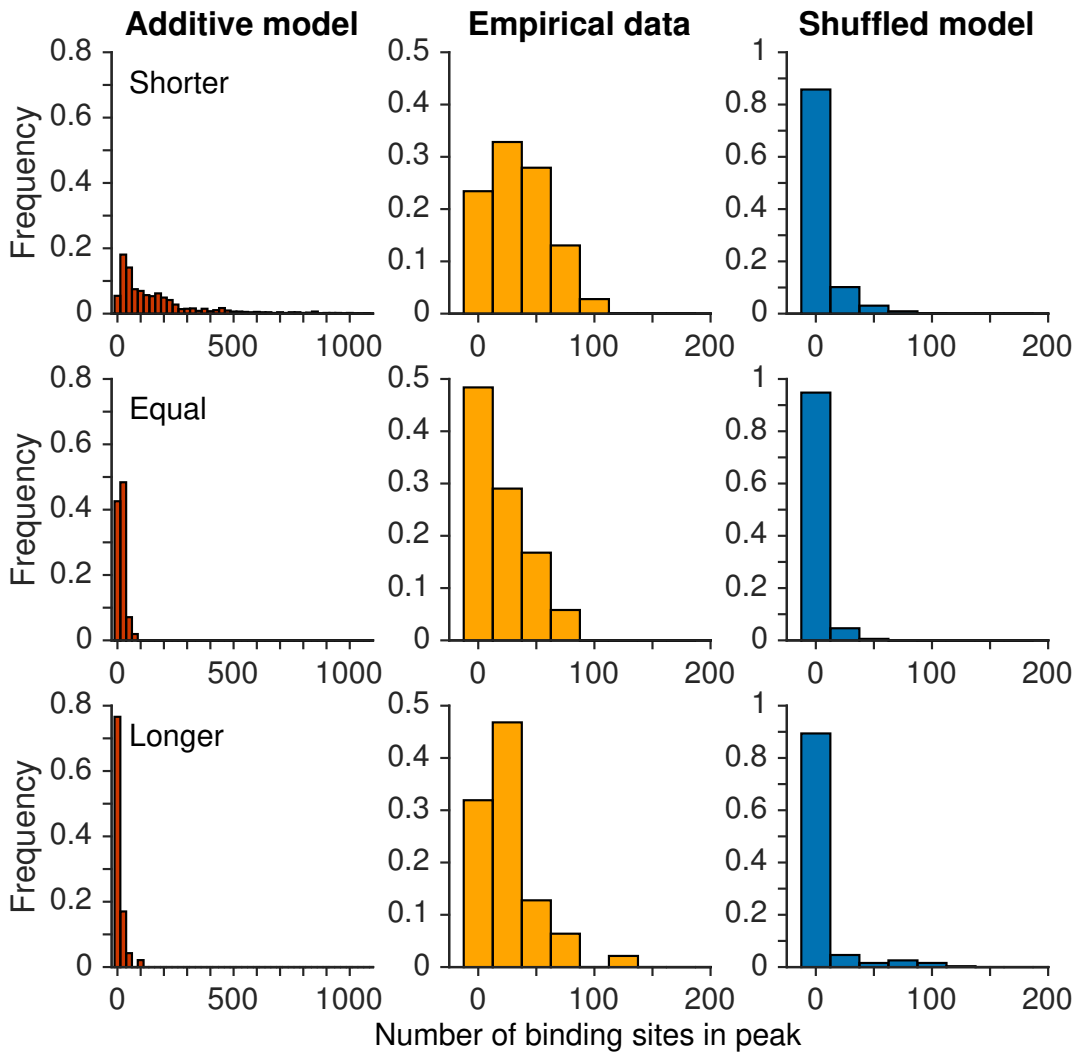
Supplementary Figure 2.22: Epistasis in the empirical data is intermediate to that of the additive and shuffled models, irrespective of DNA binding domain. These data represent a sensitivity analysis of the trends presented in Fig. 2.2b. Boxplots of magnitude epistasis, simple sign epistasis, and reciprocal sign epistasis for TFs from the 10 most prominent DNA binding domains in our dataset. The number of TFs in our dataset per DNA binding domain is shown in parentheses. Note the logarithmic scale of the y axes, which obscures data points with zero epistasis. The absence of the thick horizontal line indicates that the median of the distribution is below the lowest value of the y axis.



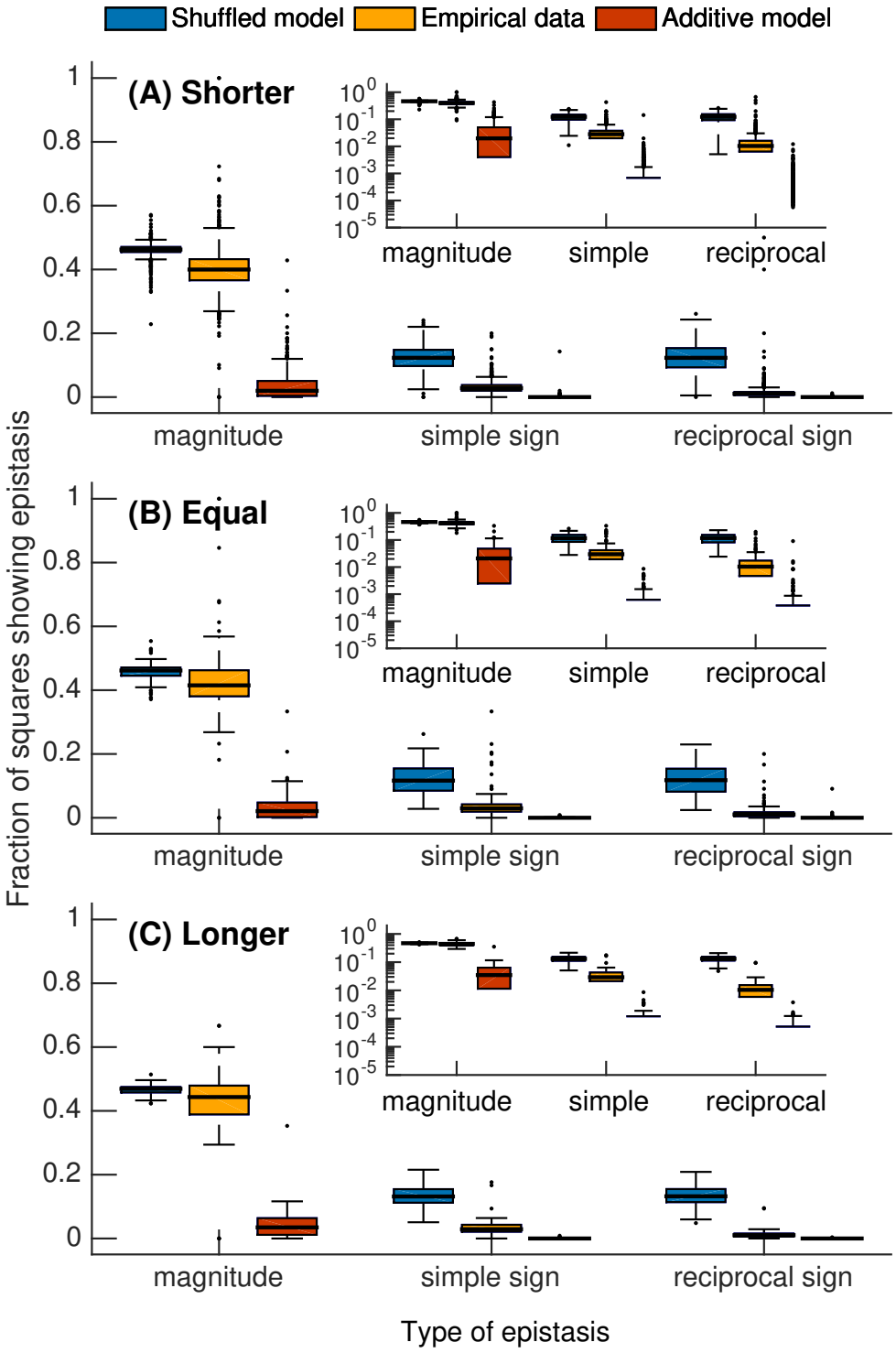
Supplementary Figure 2.23: Peak accessibility in the empirical data is intermediate to that of the additive and shuffled models, irrespective of DNA binding domain. These data represent a sensitivity analysis of the trends presented in Fig. 2.2c. Data points depict the mean peak accessibility in the empirical data and the two null models (see legend), as a function of the mutational distance, for the 10 most prominent DNA binding domains in our dataset. The number of TFs in our dataset per DNA binding domain is shown in parentheses. Error bars depict one standard deviation.



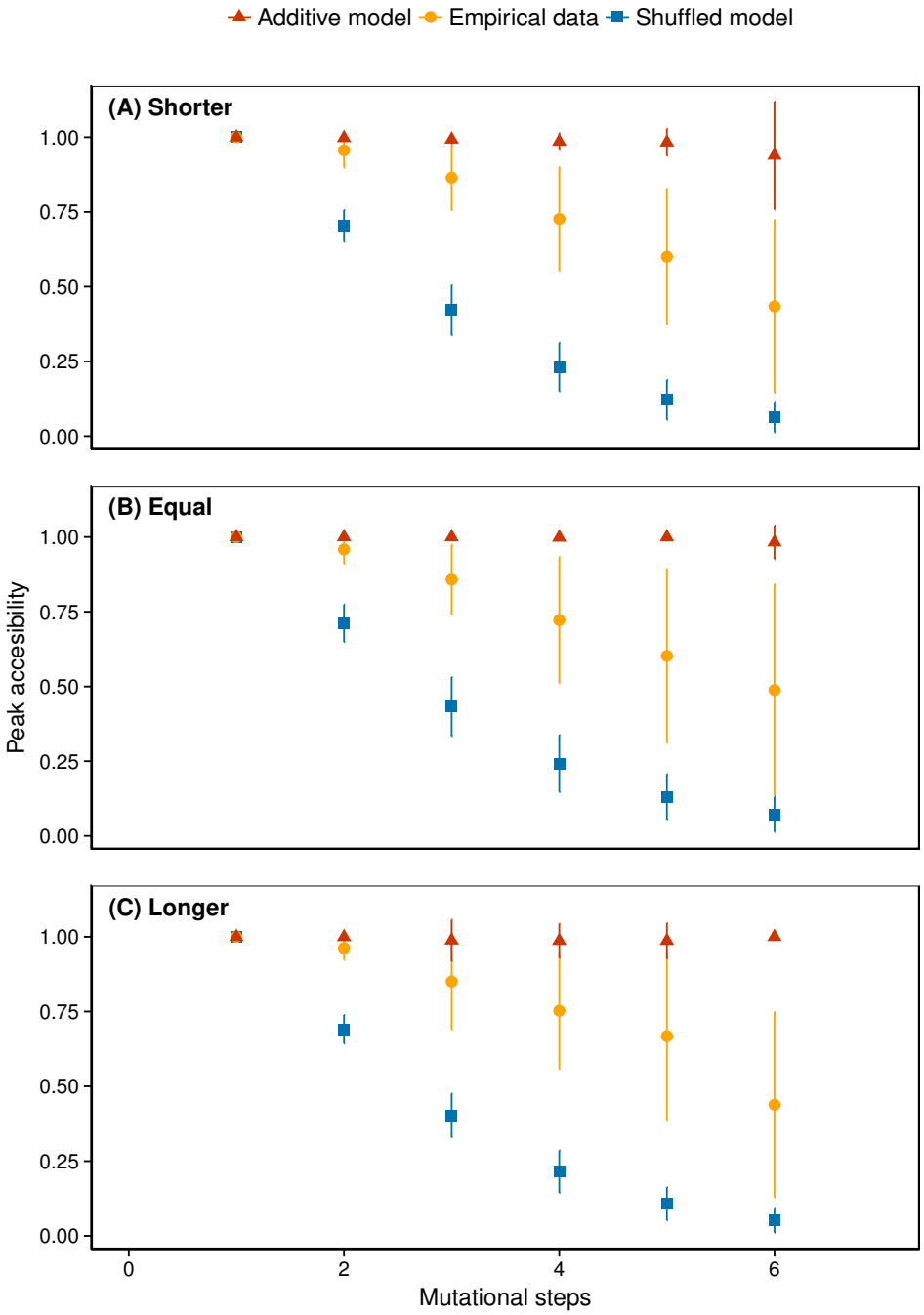
Supplementary Figure 2.24: The number of peaks in the empirical data is intermediate to that of the additive and shuffled models for TFs that bind sequences that are shorter or longer than eight nucleotides. These data represent a sensitivity analysis of the trends presented in Fig. 2.2a. Each panel shows the distribution of the number of peaks per landscape for the additive model (left column), empirical data (middle column), and shuffled model (right column). Each row of panels corresponds to TFs that bind sequences that are shorter, equal to, or longer than 8 nucleotides.



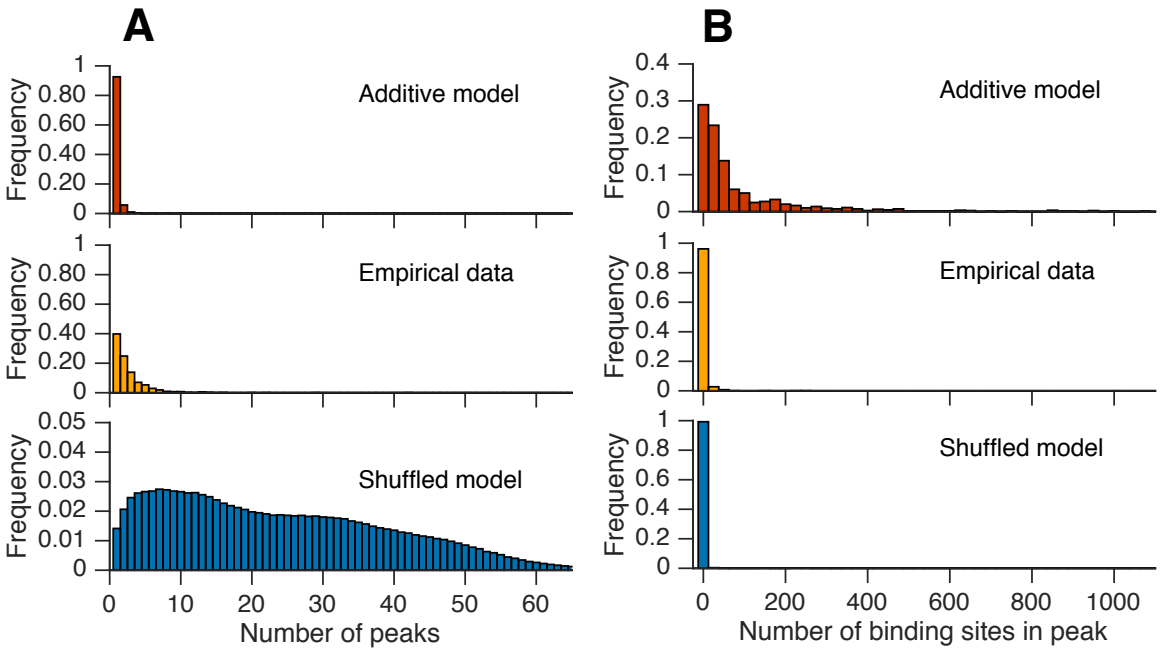
Supplementary Figure 2.25: The global peak breadth in the empirical data is intermediate to that of the additive and shuffled models for TFs that bind sequences that are shorter or longer than eight nucleotides. These data represent a sensitivity analysis of the trends presented in Supplementary Fig. 2.7. Each panel shows the distribution of the number of binding sites in the highest peak (i.e. global peak breadth) for the additive model (left column), empirical data (middle column), and shuffled model (right column). Each row of panels corresponds to TFs that bind sequences that are shorter, equal to, or longer than 8 nucleotides.



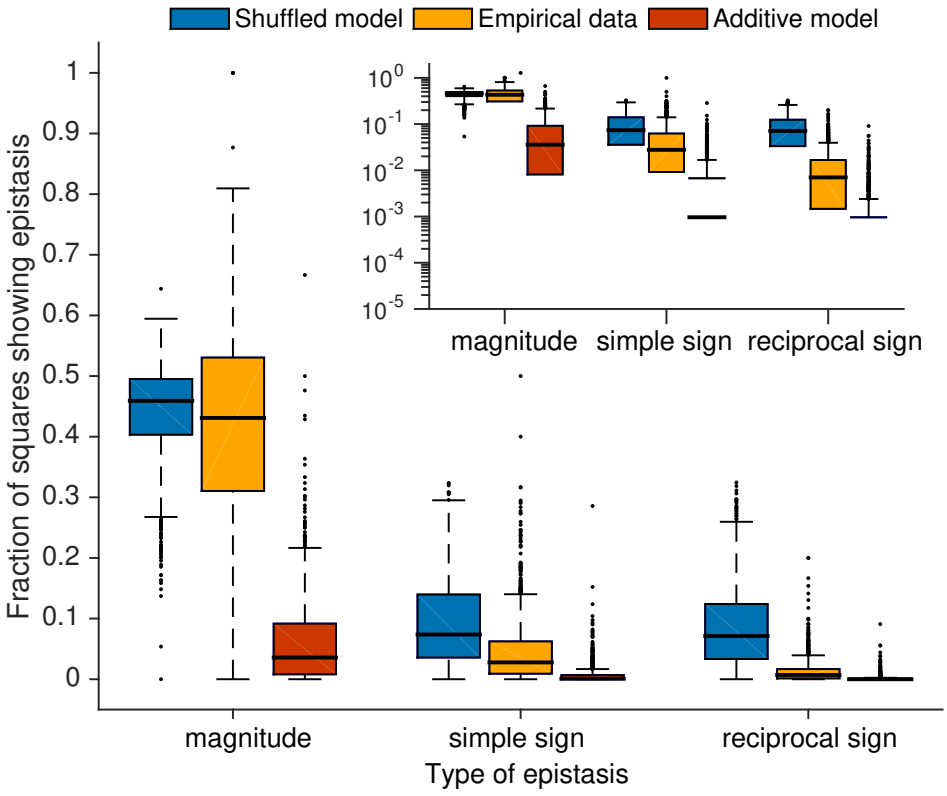
Supplementary Figure 2.26: Epistasis in the empirical data is intermediate to that of the additive and shuffled models for TFs that bind sequences that are shorter or longer than eight nucleotides. These data represent a sensitivity analysis of the trends presented in Fig. 2.2b. Boxplots of magnitude epistasis, simple sign epistasis, and reciprocal sign epistasis for TFs that bind sequences that are (**A**) shorter, (**B**) equal to, or (**C**) longer than 8 nucleotides. Note the logarithmic scale of the y axes, which obscures data points with zero epistasis. The absence of the thick horizontal line indicates that the median of the distribution is below the lowest value of the y axis.



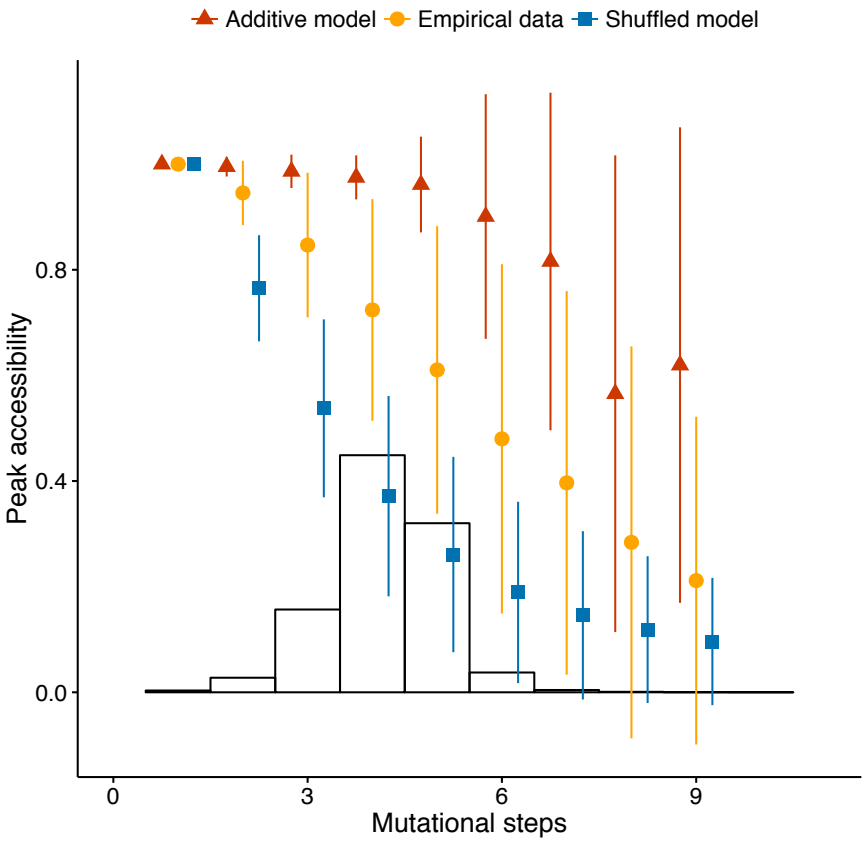
Supplementary Figure 2.27: Peak accessibility in the empirical data is intermediate to that of the additive and shuffled models for TFs that bind sequences that are shorter or longer than eight nucleotides. These data represent a sensitivity analysis of the trends presented in Fig. 2.2c. Data points depict the mean peak accessibility in the empirical data and the two null models (see legend), as a function of the mutational distance to the peak sequence, for TFs that bind sequences that are (**A**) shorter, (**B**) equal to, or (**C**) longer than 8 nucleotides. Error bars depict one standard deviation.



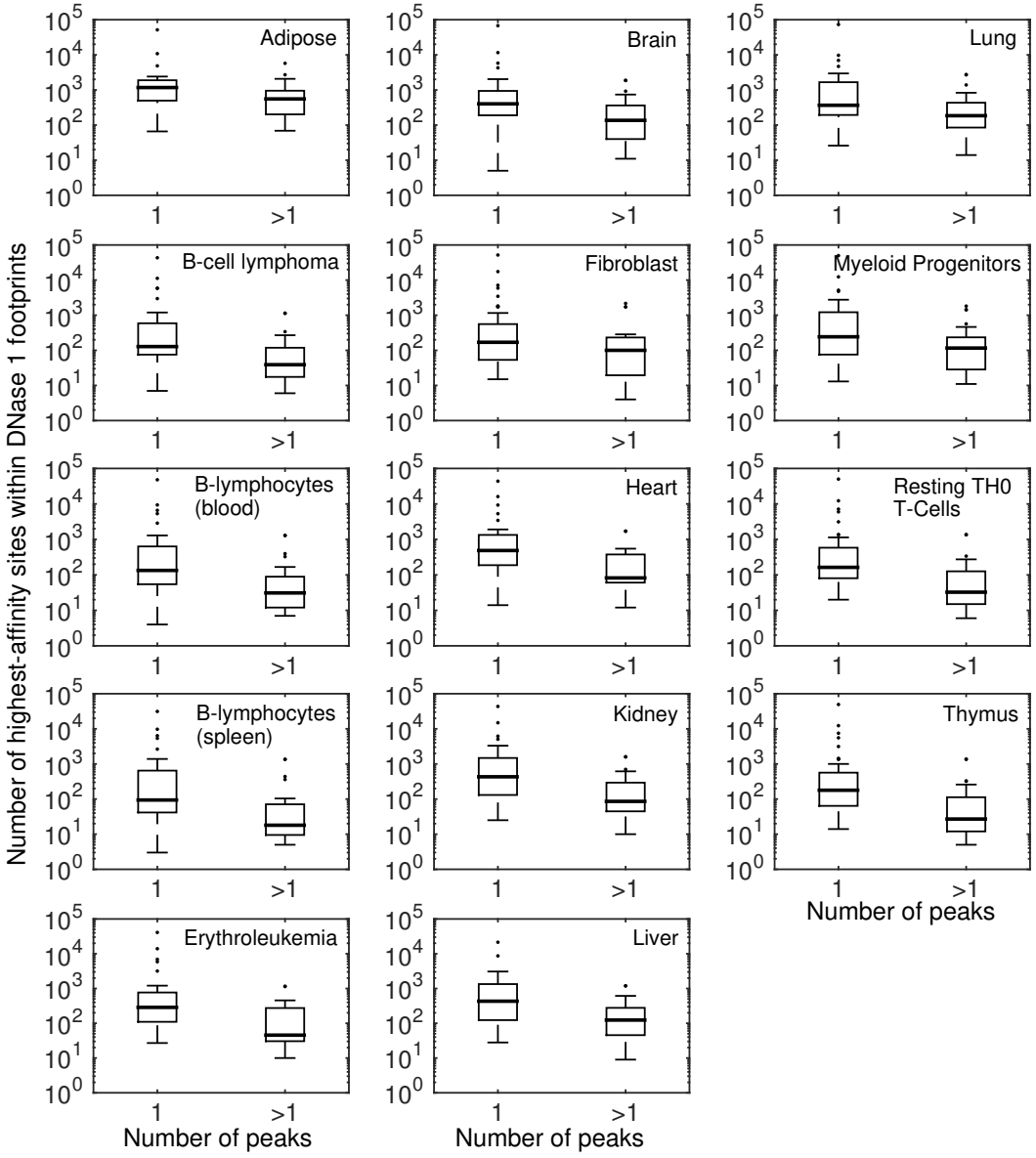
Supplementary Figure 2.28: The number of peaks in the empirical data is intermediate to that of the additive and shuffled models when using Z-scores as a proxy for binding affinity. The distribution of (**A**) the number of peaks, and (**B**) the number of binding sites per global peak for 1,095 empirical adaptive landscapes (yellow) and the additive (red) and rugged (blue) null models. These data represent a sensitivity analysis of the trends presented in Fig. 2.2c and Supplementary Fig. 2.7, respectively.



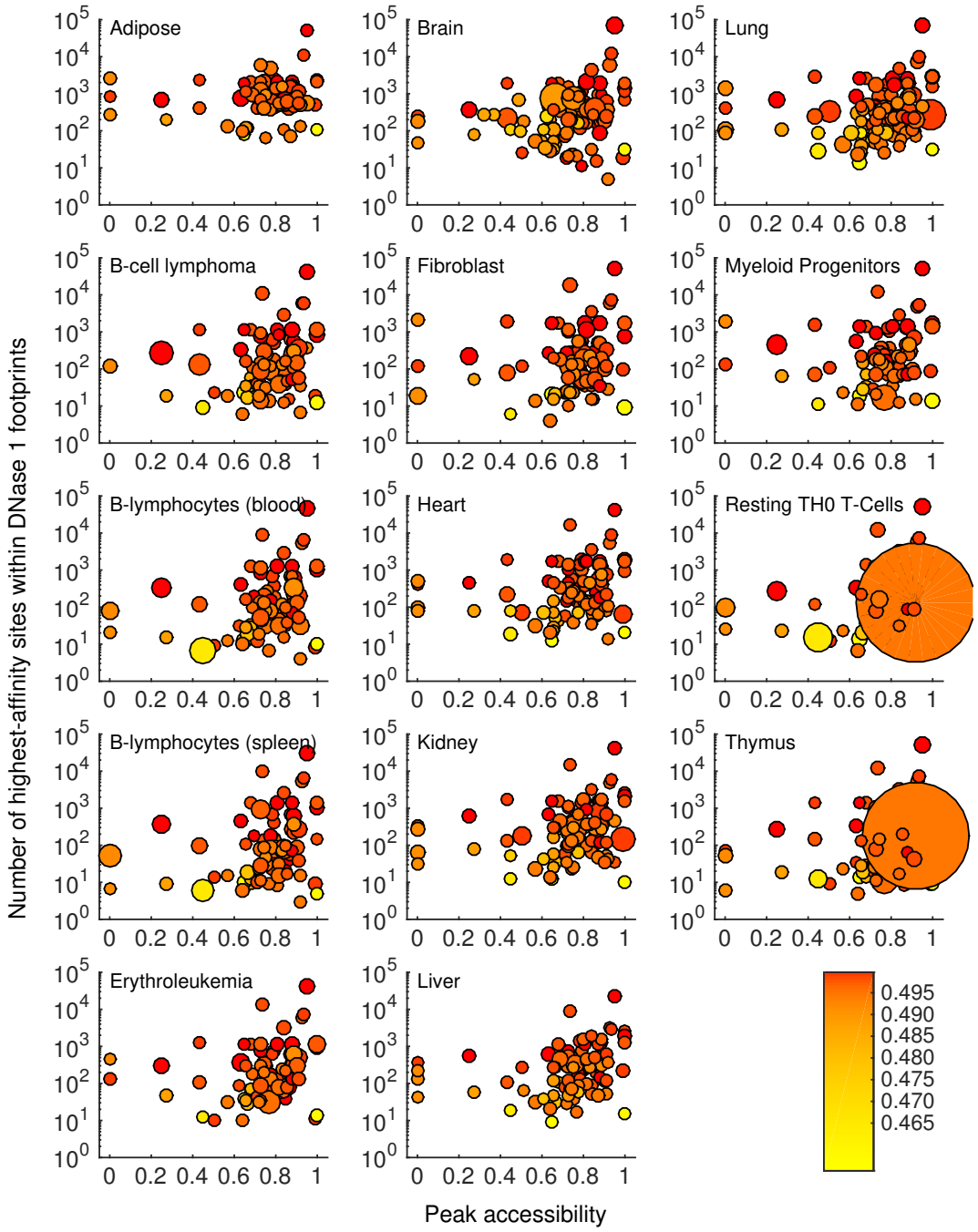
Supplementary Figure 2.29: Epistasis in adaptive landscapes of transcription factor binding affinity when using Z-scores as a proxy for binding affinity. These data represent a sensitivity analysis of the trends presented in Fig. 2.2b. Boxplots of the fraction of squares showing magnitude, simple sign, and reciprocal sign epistasis for 1,095 adaptive landscapes. The thick horizontal line in the middle of each box represents the median of the data, while the bottom and top of each box represent the 25th and 75th percentiles, respectively. For the shuffled model, the boxplot summarizes 1,095 data points, each of which is an average over 1,000 shuffled landscapes. The inset shows the same data, but with a logarithmically scaled y axis. The absence of the thick line indicates that the median of the distribution is below the lowest value of the logarithmically scaled y axis.



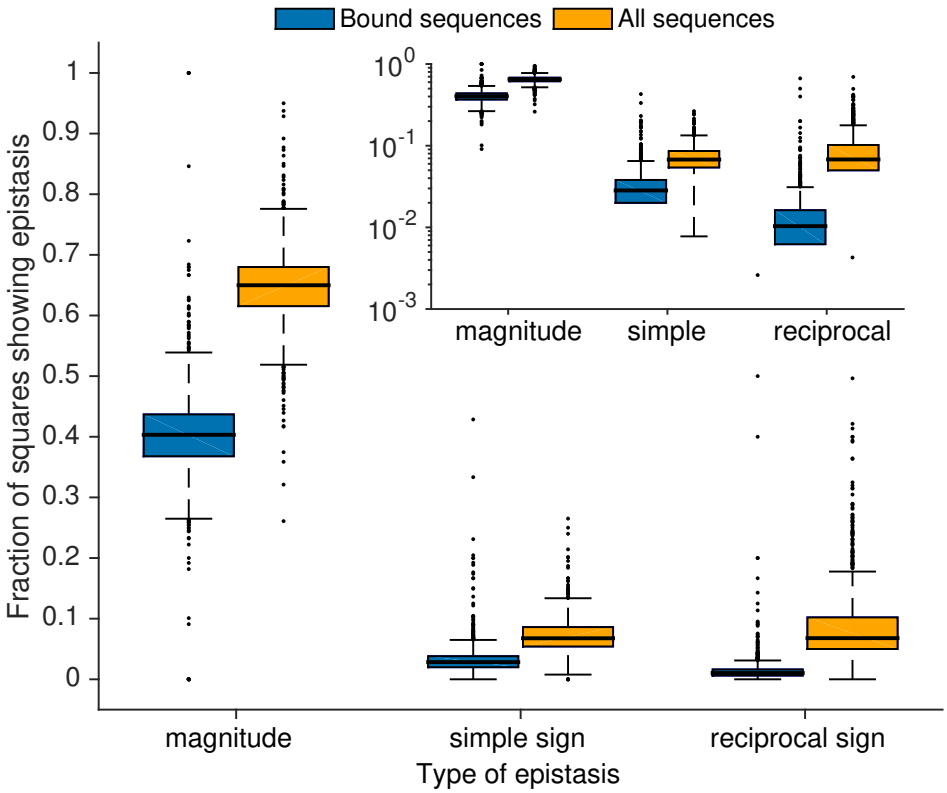
Supplementary Figure 2.30: Peak accessibility when using Z-scores as proxy for binding affinity. These data represent a sensitivity analysis of the trends presented in Fig. 2.2c. For each of the 1,095 adaptive landscapes we show the mean (symbols) and standard deviation (error bars) of the fraction of accessible paths to the highest-affinity site in the landscape (i.e., peak accessibility). The histogram shows the distribution of mutational steps from the highest-affinity site for all sequences in all landscapes. For the shuffled model, each symbol represents the mean of 1,095 data points, each of which is an average over 1,000 shuffled landscapes.



Supplementary Figure 2.31: *In vivo* **binding site abundance is higher for TFs with single-peaked landscapes than for TFs with multi-peaked landscapes.** The vertical axis of each panel indicates the abundance of a TF's highest-affinity site in protein-bound regions of the *M. musculus* genome in 14 cell and tissue types (according to DNase I footprint data; 'Methods'). The horizontal axis indicates the number of peaks. Landscapes are classified into two categories: single-peaked and multi-peaked. The thick horizontal line in the middle of each box represents the median of the data, while the bottom and top of each box represent the 25th and 75th percentiles, respectively. Each panel corresponds to a different cell or tissue type. Note the logarithmic scale of the y axis for all the panels.



Supplementary Figure 2.32: *In vivo* **binding site abundance correlates with peak accessibility.** The vertical axis of each panel indicates the abundance of a TF's highest-affinity site in protein-bound regions of the *M. musculus* genome in 14 cell and tissue types (according to DNase I footprint data; 'Methods'). The horizontal axes show peak accessibility through mutational paths of length 4, which are the most abundant paths in our dataset (Fig. 2.2c). Each panel corresponds to a different cell and tissue type, and each circle corresponds to a single TF that is expressed in that cell or tissue type. The colour of each circle indicates the binding affinity of the TF's highest-affinity site (darker = higher; colour bar). The size of a circle corresponds to the TF's expression level (larger = higher, Materials and Methods). Note the logarithmic scale of the y axis for all the panels.



Supplementary Figure 2.33: Our measures of epistasis based only on bound sequences are conservative. Box plots of the fraction of squares showing magnitude, simple sign, and reciprocal sign epistasis for 1,137 adaptive landscapes. The thick horizontal line in the middle of each box represents the median of the data, while the bottom and top of each box represent the 25th and 75th percentiles, respectively. The inset shows the same data, but with a logarithmically scaled y axis. Bound sequences are binding sites whose binding affinity is above an affinity threshold τ of 0.35 ('Methods').

3 The anatomy of an empirical genotype-phenotype map

Submitted as:

Aguilar-Rodríguez J, Peel, L, Stella, M Payne J L, and Wagner A (2017)

Abstract

Recent advances in high-throughput technologies are bringing the study of empirical genotype-phenotype (GP) maps to the fore. Here, we use data from protein binding microarrays to study an empirical GP map of transcription factor (TF) binding preferences. In this map, genotypes are DNA sequences and phenotypes are the TFs that bind these sequences. We study this GP map using genotype networks, in which nodes represent genotypes with the same phenotype, and edges connect nodes if their genotypes differ by a single small mutation. We describe the structure and arrangement of genotype networks within the space of all possible binding sites for 525 TFs from three eukaryotic species encompassing three kingdoms of life (animal, plant, and fungi). We thus provide a high-resolution depiction of the architecture of an empirical GP map. Among a number of findings, we show that these genotype networks are "small-world," assortative, can be partitioned in multiple meaningful ways, and ubiquitously overlap and interface with one another. We discuss our findings in the context of regulatory evolution.

3.1 Introduction

Evolution can be abstracted as an exploration of genotype space — the space of all possible genotypes [41]. This space is populated by intersecting sets of genotypes that each correspond to a distinct phenotype. The organization of genotype space into such geno-

type sets is described by the genotype-phenotype (GP) map [35, 42], an object of central importance in the developmental and evolutionary sciences, with important implications for medicine [12, 37, 373].

Most of what we know about GP maps comes from computational models of biological systems [59–62, 66, 155, 374]. These include models that map RNA sequence genotypes onto secondary structure phenotypes [60, 153], simplified amino acid sequence genotypes onto lattice-based, structural phenotypes [59, 375], regulatory circuit genotypes onto gene expression phenotypes [61], and metabolic genotypes onto nutrient utilization phenotypes [62]. GP maps have also been studied in non-biological systems, including self-replicating computer programs [39], evolutionary algorithms [376], and field programmable gate arrays [38]. Despite all that differentiates these systems, their GP maps have much in common. First, they are many-to-one, meaning that multiple genotypes have the same phenotype. Second, the distribution of genotypes per phenotype is heavily skewed, such that most phenotypes are realized by few genotypes, and a few phenotypes are realized by many genotypes. Third, genotypes with the same phenotype tend to be mutationally interconnected, meaning that it is possible to transform any one of these genotypes into any other via a series of small mutations that preserve the phenotype. Such sets of mutationally interconnected genotypes are known as genotype networks (aka neutral networks [60]). Fourth, the genotype networks of different phenotypes tend to overlap and interface with one another [69, 78, 286, 377]. We refer to the comprehensive description of the structure and arrangement of genotype networks within genotype space as the architecture of a GP map [378].

The architecture of a GP map has important implications for evolution, influencing the rate of adaptation [67, 68], the accumulation of genetic diversity [50], the "findability" of genotypes and phenotypes in evolutionary searches [70–72], as well as their robustness and evolvability [69]. It is therefore important to move beyond the study of GP maps derived from computational models, and to begin to study the architecture of GP maps that are derived from experimental data.

We currently know very little about the architecture of such empirical GP maps. The reason is that the genotype spaces of most biological systems are so large that it is not possible to experimentally assay a phenotype for all possible genotypes [379]. This is especially problematic when studying the architecture of a GP map, where it is necessary to assay a large number of phenotypes. Recent advances in high-throughput sequencing and chip-based technologies are beginning to mitigate this problem by providing localized descriptions of GP maps for macromolecules such as RNA and proteins [25–34]. While extraordinarily insightful, these empirical GP maps still only describe a small subset of the genotype networks of a small number of phenotypes, and therefore cannot be used to characterize the architecture of a GP map.

In contrast, protein binding microarrays [184] provide comprehensive descriptions of transcription factor (TF) binding preferences to all possible, short DNA sequences (eight nucleotides in length), and such data are available for a large number of TFs [196]. These data can therefore be used to describe the architecture of an empirical GP map, in which genotypes are DNA sequences (TF binding sites) and phenotypes are the TFs that bind these sequences. These are biologically important phenotypes, because TF binding is integral to the transcriptional regulation of gene expression, which underlies fundamental developmental, behavioral, and physiological processes in species as different as bacteria and humans [157]. What is more, DNA mutations that affect transcriptional regulation, including those in TF binding sites, may lead to evolutionary adaptations and innovations [166, 315]. Examples include binding site mutations that affect body plans in snakes [380] and the discrimination of optical stimuli in fruit flies [333].

We have recently used protein binding microarray data to characterize the topologies and topographies of genotype networks of TF binding sites [74, 156]. Our goals were to study the relationship between robustness and evolvability in TF binding sites [156], and to understand how mutation and natural selection might navigate such networks toward high-affinity binding sites [74]. To accomplish these goals, we constructed and studied the genotype networks of TFs individually, providing localized characterizations of

genotype space. Here, we extend our earlier work by providing a global and more detailed characterization of this genotype space for hundreds of TFs across three kingdoms of life, thus describing the architecture of an empirical GP map at high resolution.

3.1.1 Data

We study data from protein binding microarrays [184], a chip-based technology that measures the *in vitro* binding preferences of a TF to all possible 32,896 double-stranded DNA sequences of length eight. We refer to such sequences as TF binding sites (or simply "sites") because we study the capacity of these sequences to bind TFs. The binding preferences of a TF are reported as a list of E-scores, one per binding site. The E-score is a non-parametric, rank-based variant of the Wilcoxon-Mann-Whitney statistic that ranges from -0.5 to 0.5. It correlates with a TF's relative dissociation constant, and is therefore used as a proxy for binding affinity [184, 195]. We use this proxy to delineate sites that are bound specifically by a TF via hydrogen bond donors and acceptors (E-score > 0.35), from unbound sites or sites bound non-specifically by a TF, for example, via its affinity for the DNA backbone (E-score ≤ 0.35). We restrict our analyses to the subset of sites that are bound specifically by at least one TF [74, 156].

We consider 525 TFs from three kingdoms of life: animal, plant, and fungi (Table 3.1, Table S1). Specifically, we downloaded E-scores from the CIS-BP database for 86 TFs from *Mus musculus*, 217 TFs from *Arabidopsis thaliana*, and 118 TFs from *Neurospora crassa* [196]. We downloaded E-scores of 104 additional *M. musculus* TFs from from the UniPROBE database [195, 322]. We chose to study these three species because they have more TFs characterized in the CIS-BP database than any other in their respective kingdoms. The TFs we study collectively represent 45 unique DNA binding domains, which can be thought of as distinct biophysical mechanisms by which TFs interact with DNA. A Venn diagram of the DNA binding domains in the three species is shown in Fig. 3.1A. In our dataset, several domains are common to all three species, whereas others are unique to one species. For example, *Homeodomain* TFs are found in all three species, but the family of *Zinc cluster* TFs is exclusive to *N. crassa*. This feature of our dataset

provides an opportunity to discern whether the architecture of a GP map is governed by the peculiarities of particular binding domains or by the commonalities of TF-DNA interactions across binding domains.

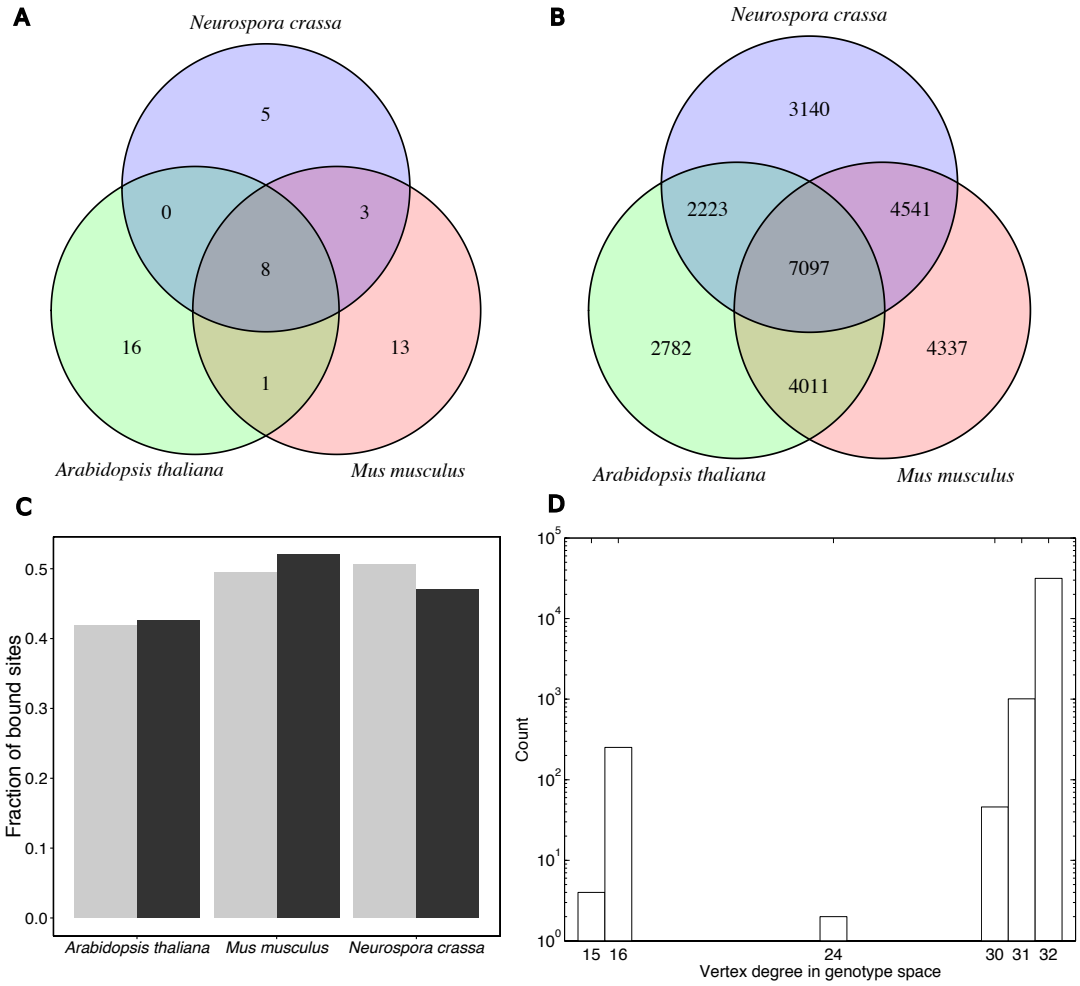


Figure 3.1: Data. (A) Venn diagram of the DNA-binding domains in the three species analyzed in this study. (B) Venn diagram of the binding repertoires of the three species. (C) Amongst all of the sites that are bound by at least one TF in a given species, the gray bars show the fraction bound by TFs with binding domains that are unique to the species, and the black bars show the fraction bound by TFs with binding domains that are not unique to the species. Bar heights do not sum to one because there are sites bound by both types of TFs. (D) Genotype space is nearly regular. Bar plot of the degree distribution of Ω. Note the logarithmic scale of the y-axis.

A Venn diagram of the sites bound by TFs from the three species is shown in Fig. 3.1B. While many sites are bound by at least one TF in all three species (21.6%), many others are bound by TFs from just a single species. Specifically, 8.5%, 13.2% and 9.6%

Table 3.1: Data analyzed in this study.

Species	Number of TFs	Number of DNA binding domains
Arabidopsis thaliana	217	25
Neurospora crassa	118	16
Mus musculus	190	25

of sites are uniquely bound by *A. thaliana*, *M. musculus*, and *N. crassa*, respectively. The TFs that bind such sites do not preferentially belong to binding domains that are exclusive to a single species (Fig. 3.1C). In total, 14.4% of the 32,896 sites are not bound by any of the TFs in our dataset.

3.1.2 Nomenclature

We consider a *genotype space* of TF binding sites for each of the three species we study. This space comprises the set of all possible 32,896 double-stranded DNA sequences of length eight. The structure of this space can be described as a network, in which nodes represent TF binding sites and edges connect nodes if their corresponding sites differ by a single small mutation, specifically by a point mutation or by an indel (Materials and Methods) [156]. We refer to this network, which contains all possible genotypes, as Ω. If two nodes are connected by an edge in Ω, we refer to them as *neighbors*.

Within this genotype space, we study a GP map in which genotypes are TF binding sites and phenotypes are the TFs that bind these sites. The set of genotypes with a particular phenotype is a *genotype set*. A single genotype may belong to multiple genotype sets, if the site binds multiple TFs. Each genotype set comprises one or more *genotype networks*, in which nodes are genotypes from the genotype set, and edges connect nodes that differ in a single small mutation, as in Ω. If a genotype set is fragmented into multiple genotype networks (connected components), it is usually the case that one network is much larger than the others [74, 156]. We refer to this network as the *dominant genotype network* (Fig. 3.2).

Genotype networks are sub-networks of Ω, in which all genotypes have the same phenotype. We refer to mutations that do not change the phenotype as *neutral*, and to

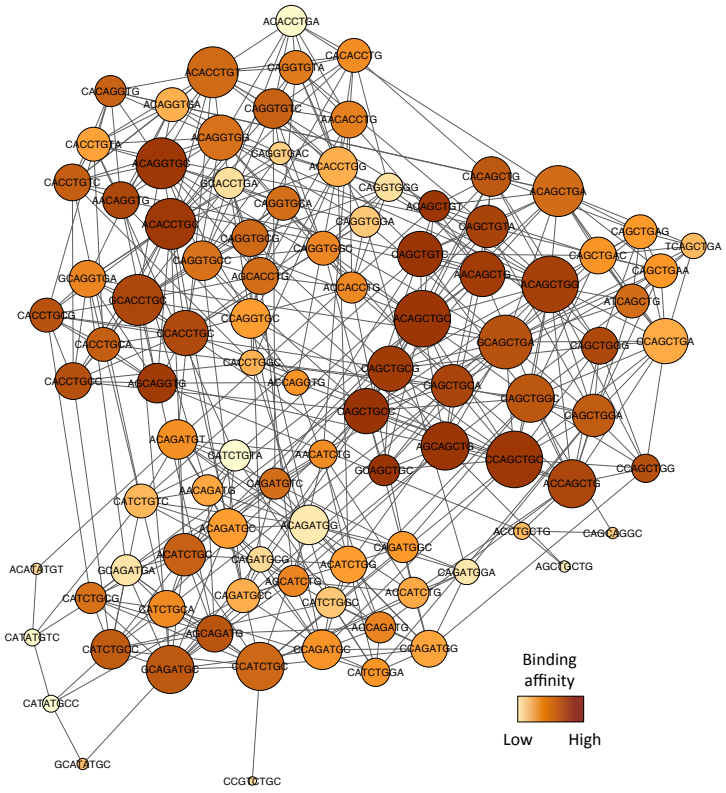


Figure 3.2: Genotype network of TF binding sites. (A) The dominant genotype network for the murine TF Ascl2. Each vertex corresponds to a DNA sequence bound by Ascl2 (E-score > 0.35). The color of a vertex indicates its binding affinity (darker = higher), while its size corresponds to the number of neighboring sequences (bigger = more). Two sequences are connected by an edge if they are separated by a single small mutation. This mutation may be a point mutation or an indel that shifts the entire binding site by a single position in either the $3'$ or $5'$ direction (Fig. 3.24).

mutations that do change the phenotype as *non-neutral*. Thus, neutral mutations define the edges within genotype networks, whereas non-neutral mutations define the edges between genotype networks, or between a genotype network and unbound sequences. If two nodes are connected by an edge in a genotype network, we refer to them as *neutral neighbors*. We emphasize that we use the term "neutral" with respect to a specific phenotype, knowing full well that such mutations may not be neutral with respect to fitness.

Non-neutral mutations bridge the genotype networks of distinct phenotypes, thus helping form the edges of a *phenotype network*. In such a network, each node represents the dominant genotype network of a specific transcription factor, and edges connect nodes if

(*i*) the associated genotype networks can be reached from one another by at least one non-neutral mutation, or (*ii*) these genotype networks share at least one genotype. In the latter case, we also say that the genotype networks *overlap*.

3.2 Results

3.2.1 Genotype space

We begin with a description of Ω, as this is the substrate of the genotype networks that we study in the subsequent sections. This network comprises 32,896 nodes and 523,728 edges. Its degree distribution of Ω is shown in Fig. 3.1D. The vast majority (96%) of genotypes have 32 neighbors, indicating that the network is nearly regular. The remaining 4% of sites possess peculiar features that are detailed in the Supplementary Material. Its diameter [381] — the longest of the shortest paths between any two nodes — is eight, which corresponds to the maximum alignment distance between two sites. On average, however, pairs of TF binding sites are separated by only 4.385 mutations. The clustering coefficient [382] of the network is 0.122, indicating that very few of a site's neighbors are neighbors themselves. The network also lacks any meaningful assortativity by degree [383] (indicated by a value of 0.006), meaning that the number of neighbors a site has provides almost no information about the number of neighbors its neighbors have. This can be attributed in part to the fact that there is very little variance in the degree distribution.

3.2.2 Intra-network analyses

3.2.2.1 General properties

We begin with some general observations about sets of genotypes bound by different transcription factors. The sizes of these genotype sets vary both within and across species, from a minimum of 2 sites for the *A. thaliana* TF Abf3 to 1,186 sites for the *M. musculus* TF Sp110. Across the three species, the average genotype set size is 374 sites. A total of 53% of these genotype sets comprise a single genotype network, whereas the remaining 47% comprise between 2 and 15 genotype networks. Despite such fragmentation, for 90%

of the TFs, more than 95% of the genotype set belongs to the dominant genotype network. We therefore carry out all of our analyses on the dominant genotype networks, as in our previous work [74, 156]. To simplify the presentation of our results, we focus on data from *M. musculus* in the main text, as it is representative of the data from *A. thaliana* and *N. crassa*, which we present in the Supplementary Material.

For the 190 *M. musculus* TFs, the average genotype network diameter is 6.7, varying from a minimum of 2 to a maximum of 14 (Fig. 3.3A). In contrast, the characteristic path length — i.e., the average shortest distance between any pair of genotypes — in a genotype network — is 3.2, less than half of the average network diameter (Fig. 3.3B). The genotype networks we study are highly clustered, with an average clustering coefficient of 0.312 (Fig. 3.3C). Taken together, the short characteristic path length relative to the diameter, and the high clustering coefficients, indicate that genotype networks of TF binding sites tend to fall within the family of "small world" networks [382]. Qualitatively similar results are obtained for the *A. thaliana* and *N. crassa* TFs (Supplementary Figs. 3.1A-C and 3.2A-C), indicating the consistency of these properties across three branches of the tree of life. The "small world" property implies that binding sites tend to be highly robust as a consequence of being highly clustered in the genotype network, while at the same time they are highly evolvable because a few mutations are enough to travel across the network and potentially access many adjacent phenotypes.

A recent numerical study suggests that a topological property known as *degree assortativity* (r) may influence evolutionary dynamics on genotype networks [68]. This measure, which ranges from $-1 \leq r \leq 1$, captures the propensity with which nodes of similar degree connect with one another (Materials and Methods) [383]. Evolutionary dynamics on genotype networks that are assortative by degree ($r > 0$) may result in *phenotypic entrapment*, where the probability that an evolving population leaves a genotype network decreases with the time spent on it [68]. We find that most genotype networks exhibit a moderate amount of degree assortativity, possessing on average a value of $r = 0.25$ (Fig. 3.3D). Degree assortativity is positively correlated with the size of the dominant geno-

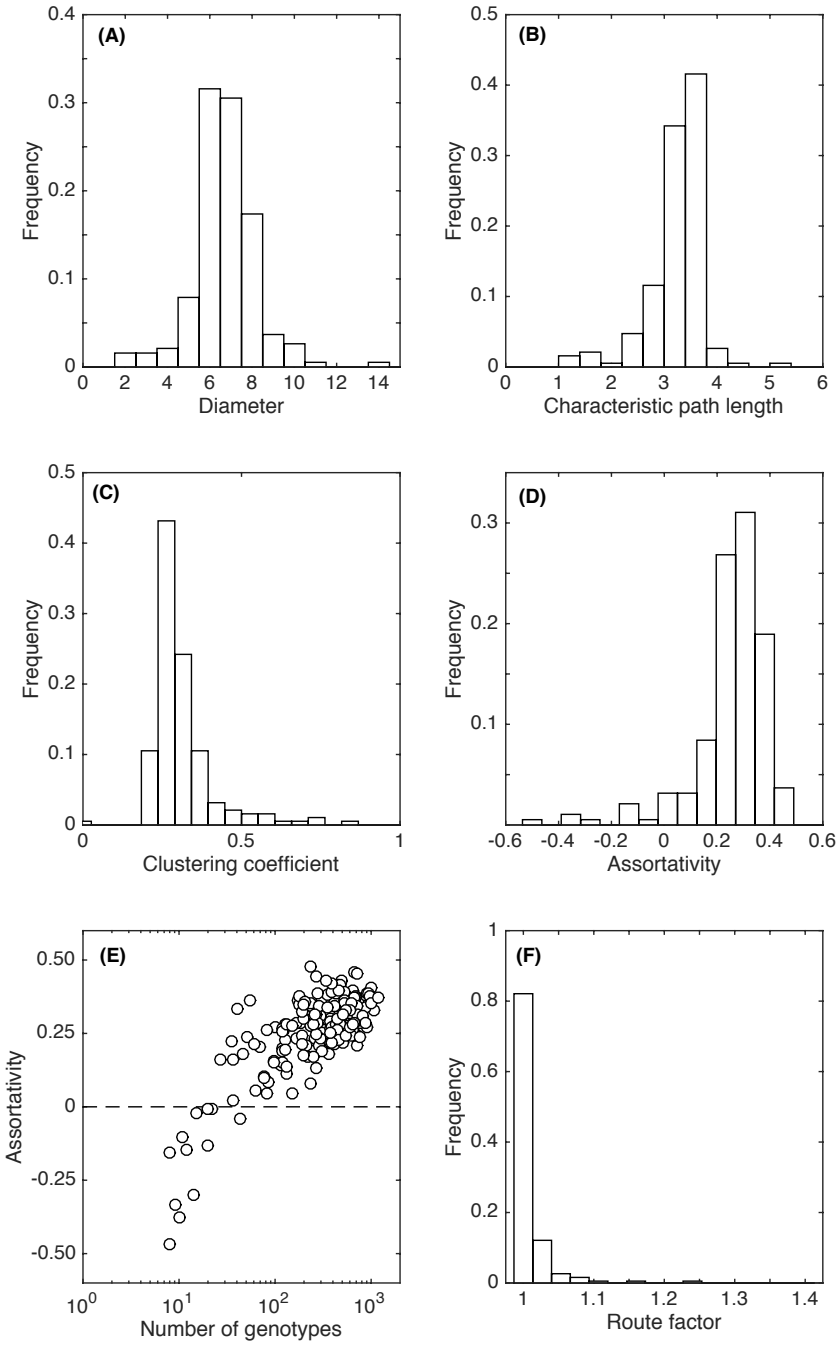


Figure 3.3: Intra-network statistics for 190 TFs from *M. musculus*. The distributions of genotype network (A) diameter, (B) characteristic path length, (C) clustering coefficient, and (D) assortativity. (E) Assortativity (horizontal axis) and its relationship to the number of genotypes in the dominant genotype network (vertical axis). The horizontal dashed line indicates an uncorrelated (non-assortative) mixing pattern. (F) The distribution of the genotype network route factor.

type network (Spearman's $r = 0.57$, $p = 1.33 \times 10^{-17}$), such that disassortative genotype networks ($r < 0$) are always small (Fig. 3.3E). This likely reflects finite-size effects. Sup-

Table 3.2: We show the number of genotype networks that have a partition that exhibits a particular group structure according to a partitioning method based on a stochastic block model.

Species	**Group structure**		
	Core-periphery	**Assortative**	**Disassortative**
A. thaliana	1 (0.46%)	213 (98.16%)	3 (1.38%)
N. crassa	1 (0.85%)	117 (99.15%)	0 (0.00%)
M. musculus	1 (0.53%)	186 (97.89%)	3 (1.58%)

plementary Figs. 3.1D,E and 3.2D,E show that the same trends also exists in *A. thaliana* and *N. crassa* TFs. Finally, we emphasize that these trends in assortativity do not simply arise from the assortativity of Ω, because Ω shows very little assortativity ($r = 0.006$).

We next describe the structure of genotype networks using a metric called the *route factor* q (Materials and Methods) [384]. It measures the average distance to a target genotype from all other genotypes in a genotype network, relative to the distance between these genotypes in Ω. When $q = 1$, the genotype network is optimally distributed in Ω, in the sense that all paths to the target genotype are the shortest possible paths. When $q > 1$, the genotype network possesses paths to the target genotype that are longer than those in Ω, indicating deviations from an optimal distribution. Fig. 3.3F shows the distribution of q for the dominant genotype networks of the 190 *M. musculus* TFs, where the target genotype is chosen to have the highest E-score. The distribution is heavily skewed toward $q = 1$, with an average route factor of $q = 1.01$. This indicates that genotype networks of TF binding sites are almost optimally distributed in Ω. Indeed, 38% of the genotype networks are optimally distributed, with $q = 1$. These results are consistent across the three species we study, as shown in Supplementary Figs. 3.1F and 3.2F.

3.2.2.2 Genotype network partitions

Networks can often be partitioned into distinct groups of nodes that have more edges within them than between them [385]. To determine if such partitions exist for genotype networks of TF binding sites, we took two approaches. In the first, we used a partitioning method that is based on a stochastic block model [385]. This method assigns each

genotype in a genotype network to one of two groups (labeled g_1 and g_2), and uses a 2×2 "mixing matrix" to describe the structure of the network. This symmetric matrix contains the probabilities of observing edges between genotypes from the same group ($p_{g_1g_1}$ and $p_{g_2g_2}$) and between different groups ($p_{g_1g_2}$). The method uses maximum likelihood to find the partition and mixing matrix that best explain the structure of the genotype network (Materials and Methods). The resulting probabilities of the mixing matrix can be used to classify each genotype network as exhibiting an assortative group structure ($p_{g_1g_1} > p_{g_1g_2} < p_{g_2g_2}$), a disassortative group structure ($p_{g_1g_1} < p_{g_1g_2} > p_{g_2g_2}$), or a core-periphery group structure ($p_{g_1g_1} > p_{g_1g_2} > p_{g_2g_2}$) [385]. We find that the vast majority of genotype networks in the mouse dataset (97.9%) exhibit an assortative group structure (Table 3.2). Thus, not only are these networks globally assortative by degree ($r > 0$, Fig. 3.3C), they are also partitionable into two groups that each have more edges within them than between them. The same is true for the *A. thaliana* and *N. crassa* TFs, of which 98% and 99% exhibit an assortative group structure, respectively (Table 3.2).

We next asked whether similar trends in group structure exist if we manually partition each genotype network according to binding affinity, rather than relying on the maximum likelihood approach described above. To do so, we used the structural partition of each genotype network into the two groups g_1 and g_2 to find an affinity threshold that best separates the binding affinities of these groups (Materials and Methods). We used this threshold to label the genotypes as belonging to a high-affinity group g_{high} or to a low-affinity group g_{low}. We then constructed a mixing matrix that contains the probabilities of observing edges within groups ($p_{g_{\text{low}},g_{\text{low}}}$ and $p_{g_{\text{high}},g_{\text{high}}}$) and between groups ($p_{g_{\text{low}},g_{\text{high}}}$), calculated directly from each genotype network. We used this mixing matrix to test the null hypothesis H_0 that binding affinity is distributed uniformly at random with respect to the structure of the genotype network (Materials and Methods) [386]. Thus, rejection of H_0 indicates that the binding affinity partition provides meaningful information about genotype network structure. Table 3.3 shows that H_0 is almost always rejected. On the rare occasion that H_0 is accepted, the genotype network is small ($\leq$ 72 nodes), which

Table 3.3: The number of genotype networks that have a binding affinity partition that exhibits a particular group structure. We also test the null hypothesis H_0 that binding affinity is random with respect to genotype network structure, rejecting H_0 if $p < 0.05$.

Species	**Group structure**			H_0
	Core-periphery	**Assortative**	**Disassortative**	**rejections**
A. thaliana	134 (61.75%)	79 (36.41%)	4 (1.84%)	209 (96.31%)
N. crassa	67 (56.78%)	49 (41.53%)	2 (1.69%)	113 (95.76%)
M. musculus	118 (62.11%)	66 (34.74%)	4 (2.11%)	181 (95.26%)

again likely indicates finite-size effects. Additionally, we find in *M. musculus* that 62.1% of the binding affinity partitions exhibit a core-periphery group structure ($p_{g_{\text{high}},g_{\text{high}}} > p_{g_{\text{high}},g_{\text{low}}} > p_{g_{\text{low}},g_{\text{low}}}$), while 34.7% exhibit an assortative group structure ($p_{g_{\text{high}},g_{\text{high}}} > p_{g_{\text{high}},g_{\text{low}}} < p_{g_{\text{low}},g_{\text{low}}}$). Similar results are obtained for the *A. thaliana* and *N. crassa* TFs (Table 3.3). In sum, genotype networks of TF binding sites can be partitioned in multiple meaningful ways, and the resulting group structure depends upon how the partition is defined. An assortative group structure is uncovered by a structure-based partition, whereas a core-periphery group structure can be uncovered by an affinity-based partition.

3.2.3 Inter-network analyses

We now shift the scale of our analysis from local to global, transitioning from descriptions of individual genotype networks to descriptions of how these genotype networks overlap and interface with one another in Ω.

3.2.3.1 Overlap

Some TFs have similar binding preferences, especially if they are products of duplicate (paralogous) genes [195, 196]. The genotype networks of such TFs will therefore overlap. Fig. 3.4A shows the extent of this overlap for all pairs (p, q) of TFs in the mouse dataset. Rows and columns correspond to individual TFs, and are arranged by DNA binding domain. The shading of matrix elements depicts overlap as the fraction of binding sites that are common to the genotype networks of two TFs. The matrix is asymmetric, because overlap is normalized by the genotype network size of TF q (Material and Methods).

Similar values of overlap are found in *A. thaliana* and *N. crassa* (Supplementary Figs. 3.3A, 3.4A).

Paralogous TFs exhibit a high level of overlap in their genotype networks, as indicated by the block structure of the main diagonal in Fig. 3.4A. Even TFs with a *C2H2 ZF* binding domain, which exhibit the lowest levels of overlap, still share 9.14% of their binding sites on average. At the other end of the spectrum are two TFs with an *E2F* binding domain (E2F2 and E2F3), which share 92.73% of their binding sites. Overlap is not restricted to TFs from the same binding domain, as indicated by the blue shading off the main diagonal. For example, *ARID/BRIGHT* and *Sox* TFs share on average 16.5% of their binding sites. In fact, every single TF in the *M. musculus* dataset exhibits overlap in its genotype network with at least one other TF from a different binding domain. The total number of TFs that overlap a TF's genotype network ranges from a minimum of 18 for the TFs Olig3 and Tcfap2e to a maximum of 162 for the TF Eomes. TFs with the same DNA binding domain tend to share on average 27.2% of their binding sites, while TFs with different binding domains only share 1.88% on average (Wilcoxon rank-sum test, $p < 10^{-6}$).

The number of binding sites in the region of overlap between two TFs ranges from a minimum of 1 for 1,730 pairs of TFs to a maximum of 884 for the TFs E2F2 and E2F3. The average number of binding sites in overlapping regions is 29.38. The size of the overlapping region is larger between members of the same binding domain than between TFs with different binding domains (Wilcoxon rank-sum test, $p < 10^{-6}$). The average size of the overlapping region between TFs with the same binding domain equals 71.46 sites, while between TFs with different binding domains it equals 7.09 sites.

We next asked whether the binding affinity of sites to one TF are correlated with their affinity to another TF, for binding sites that are contained in both genotype networks. We studied $3,467$ overlapping regions between pairs of TFs that contain more than 10 binding sites. We found that in 79% of the overlapping regions, binding affinities are not correlated (FDR-adjusted $p > 0.05$). Among the regions that show a significant correlation (FDR-

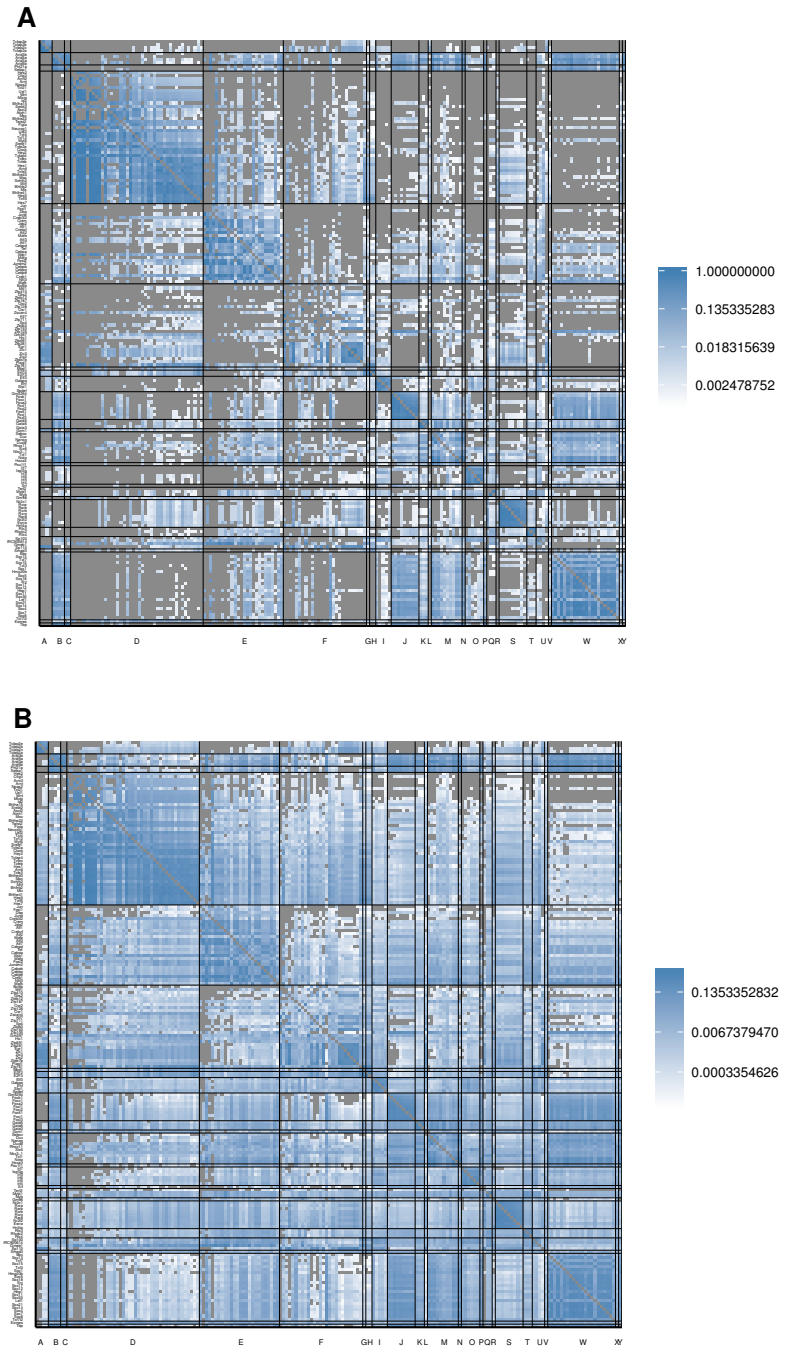


Figure 3.4: Matrices of inter-network relationships for the genotype networks of TF binding sites from *M. musculus*. Heatmaps of log10-transformed (A) overlap and (B) ϕ_{qp}, the probability of mutating from the genotype network of phenotype p to the genotype network of phenotype q. The rows and columns are grouped according to binding domain, which are ordered alphabetically on the horizontal axis: A, AP-2; B, ARID/BRIGHT; C, AT hook; D, bHLH; E, bZIP; F, C2H2 ZF; G, CxxC; H, E2F; I, Ets; J, Forkhead; K, GATA; L, GCM; M, Homeodomain; N, Homeodomain + POU; O, IRF; P, MADS box; Q, Myb/SANT; R, Ndt80/PhoG; S, Nuclear receptor; T, RFX; U, SAND; V, SMAD; W, Sox; X, T-Box; Y: TBP. Within the DNA-binding domain groups, the rows and columns are ordered by the size of each TF's dominant genotype network, such that network size increases from top to bottom and from left to right. Labels on the vertical axis indicate the name of the TFs. Cells colored in gray indicate either N/A values (on the diagonal) or values equal to zero (off-diagonal).

adjusted $p < 0.05$), 93% of them show a positive correlation (Spearman's $r > 0$). This is especially the case for TFs that have the same binding domain, where the average correlation coefficient is 0.536. In contrast, for TFs with different binding domains the average correlation coefficient is 0.094 (Wilcoxon rank-sum test, $p < 10^{-6}$).

We then studied if there is a relationship between the genotype network partitions and the regions of overlap. Is any of the two groups of nodes (g_1 or g_2) into which a network is partitioned over-represented in the overlapping regions between pairs of networks? To investigate this question, we studied the fraction of the overlap between networks i and j that fall within groups g_1 or g_2 into which network i is partitioned (group g_1 has always a higher mean degree than g_2). We compared the observed fraction to a null distribution of fractions obtained via 1000 permutations of the assignment of g_1 or g_2 to the binding sites in network i. The number of permuted networks in which this fraction is smaller than that of the measured fraction can be used to calculate an empirical p-value for each i, j pair. We first analyzed the relationship between the regions of overlap and the partitions obtained according to the stochastic block model. Of the 3,467 pairs, 59.8% have an over-representation of either g_1 or g_2 (FDR-adjusted $p < 0.05$). Of that percentage, 54.9% of pairs show an over-representation of partition g_1, and 45.1% of partition g_2. Thus, we do not find a significant relationship between overlap and the stochastic block model partitions for all pairs. In contrast, we do find that the fraction of overlapping regions where the partition g_1 is over-represented is higher in pairs of TFs with the same binding domain (71.64%) than in pairs of TFs with different binding domains (41.88%) (Fisher's exact test: odds ratio $F = 3.51$, $p < 10^{-6}$).

We then repeated the above analysis using the binding affinity partitions. Only 34% of overlapping regions exhibit an over-representation of a partition. Of that percentage, 93.6% of overlapping regions show an over-representation of the high-affinity partition g_{high}, while the low-affinity partition g_{low} is only over-represented in 6.4% of pairs. We also find that the fraction of overlapping regions where the partition g_{high} is over-represented is higher in pairs of TFs with the same binding domain (98.75%) than in pairs of TFs with

different binding domains (84.22%) (Fisher's exact test: odds ratio $F = 14.85$, $p < 10^{-6}$).

3.2.3.2 Interface

To characterize how the genotype networks of TF binding sites interface with one another, we calculate the fraction ϕ_{qp} of mutations to binding sites in the genotype network of TF p that create binding sites in the genotype network of TF q (Materials and Methods) [122]. The matrix in Fig. 3.4B shows ϕ_{qp} for all TFs in the mouse dataset. It is arranged as in Fig. 3.4A. Similar values of ϕ_{qp} are found in *A. thaliana* and *N. crassa* (Supplementary Figs. 3.3B, 3.4B).

Of the 35,910 pairwise comparisons depicted in Fig. 3.4B, 31,548 (87.9%) have $\phi_{qp} > 0$. This means that genotype networks of TF binding sites interface with one another to such an extent that it is usually possible to evolve at least one of a TF's binding sites via a single small mutation to a binding site of nearly any other TF. Non-zero ϕ_{qp} values range from a minimum of 2.63×10^{-5} from the TF Sp110 to the TFs Phf21a, Npas2 and Foxb1, to a maximum of 0.84 from the TF Arnt2 to the TF Bhlhe41. On average, the ϕ_{qp} between the genotype networks of TFs with the same binding domain is higher than that of TFs with different binding domains (0.139 compared to 0.016; Wilcoxon rank-sum test, $p < 10^{-6}$), but there are some exceptions. For example, the genotype networks of TFs with a *SAND* binding domain have a higher ϕ_{qp}, on average, with the genotype networks of TFs with a *bZIP* binding domain than they do with the genotype networks of TFs with the same binding domain. To investigate this further, we compare ϕ_{qp} to an expected value that is generated using a null model (Materials and Methods) [156]. This expected value is equivalent to the fraction f_q of genotypes with phenotype q [72, 122]. We find that the null model provides a reasonable approximation to the empirical data (Fig. 3.5), echoing earlier observations in computational models of GP maps [122]. This means that the overall frequency of a phenotype — i.e., the fraction of genotypes with that phenotype — is a good indicator of the probability that a randomly chosen non-neutral mutation leads to that phenotype. When the quantity ϕ_{qp} is larger than the null expectation (i.e., above the solid line in Fig. 3.5), it is often the case that TFs p and q have the same

binding domain (Fig. 3.5, filled circles). Since such TFs often bind similar sets of sites [196], this observation corroborates the intuition that their genotype networks interface more than expected by chance. Across all pairs of TFs with the same binding domain, ϕ_{qp} is greater than the null expectation for 84.7% of pairs. In contrast, across all pairs of TFs with different binding domains, ϕ_{qp} is greater than the null expectation for only 28.1% of pairs.

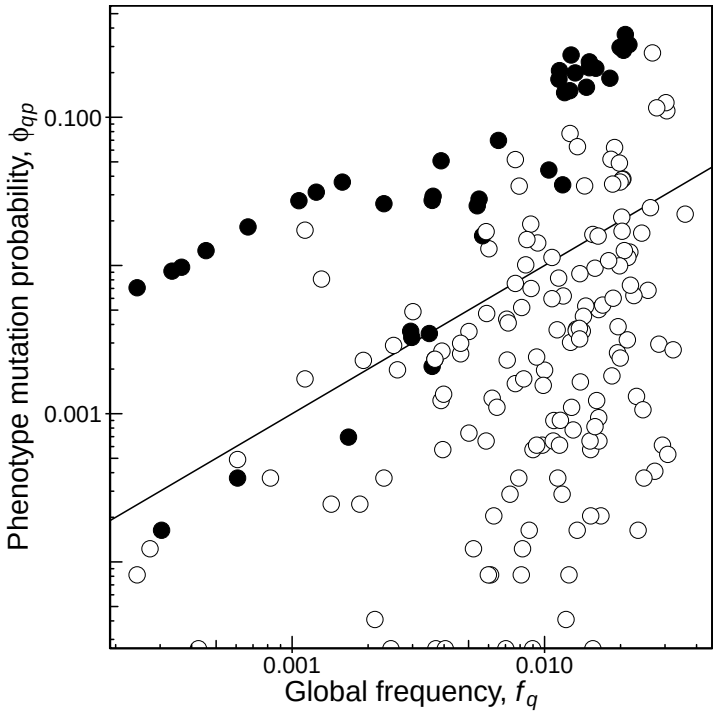


Figure 3.5: A simple null model provides a reasonable approximation to ϕ_{qp}. The probability ϕ_{qp} that a mutation to a genotype with phenotype p creates a genotype with phenotype q is shown in relation to the frequency f_q of phenotype q. The black line shows the null expectation that $\phi_{qp} = f_q$ [122]. Each circle represents the ϕ_{qp} of a different phenotype p, where phenotype q is always that of murine TF Sp110. Black circles correspond to TFs with the same binding domain as Sp110 (i.e., SAND), and white circles correspond to TFs with a different binding domain. Half circles at the bottom of the panel denote pairs of phenotypes with $\phi_{qp} = 0$.

While ϕ_{qp} is well approximated by the null model, it is also a quantity that is averaged across an entire genotype network. Since GP maps often exhibit correlations in their local mutational neighborhoods [122], we sought to determine if the composition of such neighborhoods — in terms of the phenotypes that occur in them — might deviate from the null expectation. To do so, we compared the composition of the mutational neighborhoods

of pairs of neighboring genotypes on a genotype network to the mutational neighborhoods of randomly selected pairs of non-neighboring genotypes from the same genotype network, removing neighbors that are shared by the genotypes being compared. We used this comparison to compute a similarity ratio that is greater than unity when neighboring genotypes have more similar sets of phenotypes in their mutational neighborhoods than do non-neighboring genotypes (Materials and Methods) [122]. Fig. 3.6 shows a histogram of this similarity ratio for all possible pairs of neighboring genotypes in the genotype network for the mouse TF Sp110, which we have chosen to exemplify this result because it has the largest genotype network in the *M. musculus* dataset. The mean is 1.465 ± 0.006, which deviates significantly from the null expectation of unity (one-sample t-test, $t = 79.87$, $p < 10^{-6}$). Across all mouse TFs, the average similarity ratio ranges from a minimum of 1.029 for the TF Usf1 to a maximum of 3.302 for the TF Zfp187. We made similar observations in the *A. thaliana* and *N. crassa* data (Supplementary Figs. 3.5, 3.6).

So far, we have only considered how genotype networks interface with one another. Since mutations that abrogate TF binding are also important for regulatory evolution [380], we now turn our attention to the interface of genotype networks with the regions of Ω that do not bind any TF. Such unbound regions are not small: They comprise 51%, 48%, and 39% of Ω in *A. thaliana*, *N. crassa*, and *M. musculus*, respectively. For each TF p, we calculate the fraction $\phi_{\text{unbound},p}$ of mutations to binding sites in the genotype network for TF p that create unbound sites — i.e., sites that do not bind any TF in our dataset, for the respective species. We then divide this number by the fraction f_{unbound} of unbound sites, which is the null expectation for $\phi_{\text{unbound},p}$ [122]. Thus, this ratio will equal unity when the empirical data is well represented by the null model. Fig. 3.7 shows this ratio for all of the mouse TFs. It is consistently below unity, ranging from 0.003 for the TF Hes2 to 0.638 for the TF Zbtb12, with an average value of 0.154. This indicates that unbound sites occur less frequently in the mutational neighborhoods of bound sites than is expected under the null model. Thus, the interface of genotype networks with unbound sites in Ω is qualitatively different from the interface of genotype networks with one another. We

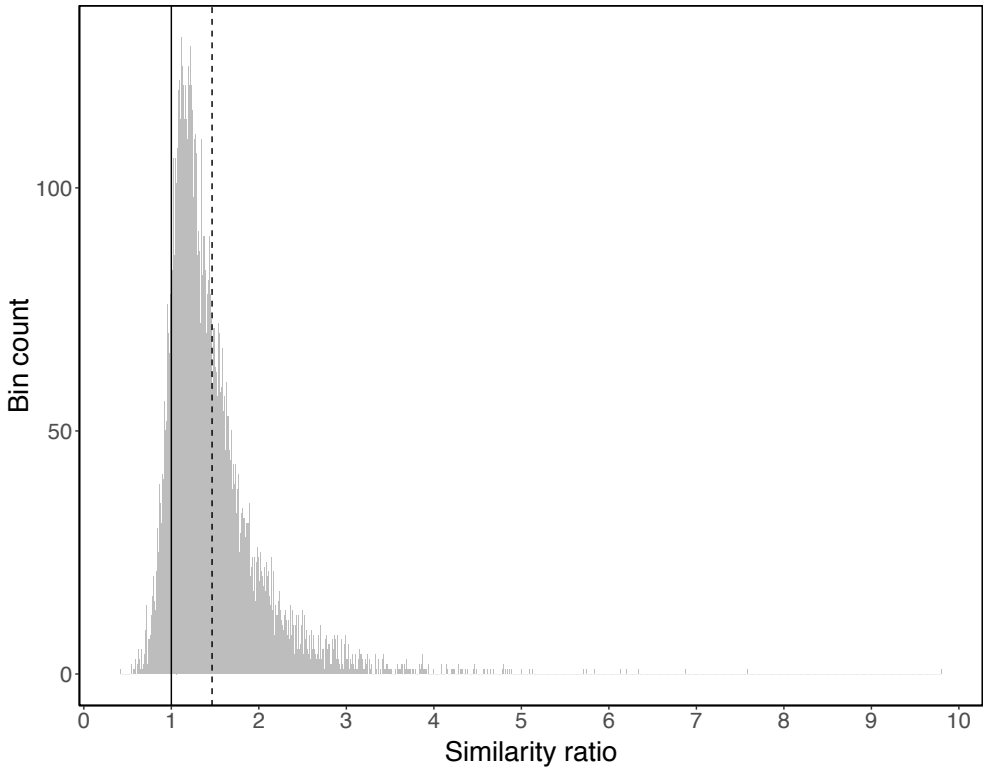


Figure 3.6: The phenotypes found in the mutational neighborhoods of neutral neighbors are more similar than those of neutral pairs that are not neighbors. The distribution of the similarity ratio (Eq. 3.12) of the phenotype probability distributions (Eq. 3.9) is shown for neutral neighbors (n_1 and n_2) and neutral pairs that are not neighbors (n_1 and n_3). For this analysis, we considered all 9,207 pairs of neutral neighbors in the genotype network for Sp110, and sampled the same number of neutral pairs that are not neighbors. The mutational neighborhoods of n_1 and n_2 are more similar than those of n_1 and n_3, because the mean of the ratio (vertical dashed line) is larger than unity (vertical solid line). The standard error of this mean (0.006) is minute compared to the difference between the mean and unity (0.465).

made similar observations in the *A. thaliana* and *N. crassa* data (Supplementary Figs. 3.7, 3.8).

Finally, we sum across the columns of Fig. 3.4B to obtain a global measure Φ_q of the mutational connectivity of the genotype network of phenotype q with the genotype networks of all other phenotypes in genotype space (Materials and Methods). This measure is related to, and highly correlated with a popular measure called *phenotypic accessibility* [70, 152] (Spearman's $r = 0.95$, $p < 10^{-6}$; Supplementary Fig. 3.9; Materials and Methods). The main difference is that Φ_q accounts for genotype network overlap. We

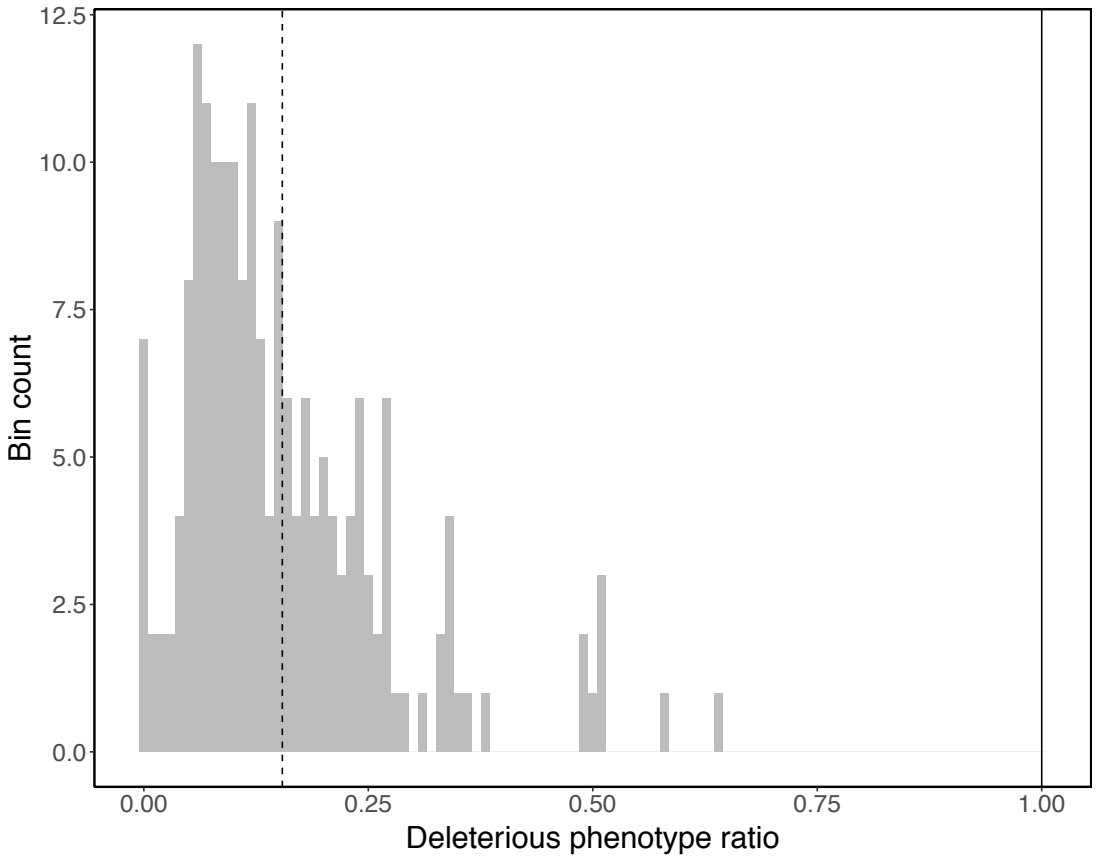


Figure 3.7: In *M. musculus*, unbound sites are underrepresented in the neighborhoods of bound sites. The distribution of the ratio $\phi_{\text{unbound},p}/f_{\text{unbound}}$, which is the probability of mutating from a sequence bound by TF p to an unbound sequence, divided by the null expectation of the frequency of unbound DNA sequences. The distribution is clearly skewed to values smaller than one, as shown by the distance of the distribution (vertical dashed line) to unity (vertical solid line).

find that Φ_q increases with genotype network size (Spearman's $r = 0.64$, $p < 10^{-6}$; Supplementary Fig. 3.10), indicating that non-neutral mutations to TF binding sites are more likely to create binding sites for low-specificity TFs than for high-specificity TFs, because low-specificity TFs have larger genotype networks [156]. We also find that Φ_q increases with genotype network size in *A. thaliana* (Supplementary Fig. 3.11) and *N. crassa* (Supplementary Fig. 3.12).

3.2.3.3 Phenotype space covering

To further characterize how genotype networks of TF binding sites overlap and interface with one another, we calculated the average fraction of phenotypes found within n

mutations of each binding site, for each TF. We refer to this measure, which has been introduced in a different context as *shape space covering* [60], as *phenotype space covering*, and we refer to a phenotype that is found within a mutational radius of a genotype as "covered." We again use the mouse TF Sp110 to exemplify our findings.

We considered two variants of phenotype space covering. In the first, we determined the phenotypes of all genotypes within a mutational radius of n, such that all mutations are neutral (i.e., the binding sites are part of the same genotype network). This analysis is therefore a further characterization of genotype network overlap. We find for the murine TF Sp110 that within just a single mutation ($n = 1$), an average of 8.51% of the phenotypes are covered, and that within a mutational radius of $n > 4$, a total of 46.31% of the phenotypes are covered (Fig. 3.8A). The genotype network for Sp110 therefore overlaps with the genotype networks of nearly half of the mouse TFs in our dataset. We then asked how the maximum proportion of phenotypes covered (e.g., 46.31% for Sp110) relates to the size of a genotype network. Fig. 3.8B shows that this maximum proportion is largely determined by the size of the dominant genotype network (Spearman's $r = 0.76$, $p < 10^{-6}$), such that larger dominant genotype networks cover more phenotypes.

In the second variant of phenotype space covering, we considered all genotypes within a mutational radius of n, such that all mutations are non-neutral. The proportion of phenotypes covered within a mutational radius of $n = 1$ does not differ from the first variant, but it increases more rapidly with n, such that all phenotypes are covered within a mutational radius of $n > 4$ (Fig. 3.8A). Moreover, there is no variation in this measure when $n > 4$, meaning that all phenotypes are covered within this mutational radius from any binding site of Sp110. Across all of the mouse TFs, $n = 4.5$ is the average mutational radius for which the coefficient of variance (σ/μ) in the proportion of phenotypes covered becomes smaller than 1%. There are 33 TFs that cover more than 99% of all phenotypes within a radius of $n \leq 4$. Remarkably, 5 of these networks are extremely small, comprising between 8 and 11 binding sites (TFs Arnt2, Fosl1, Hes2, Jun, and Olig3).

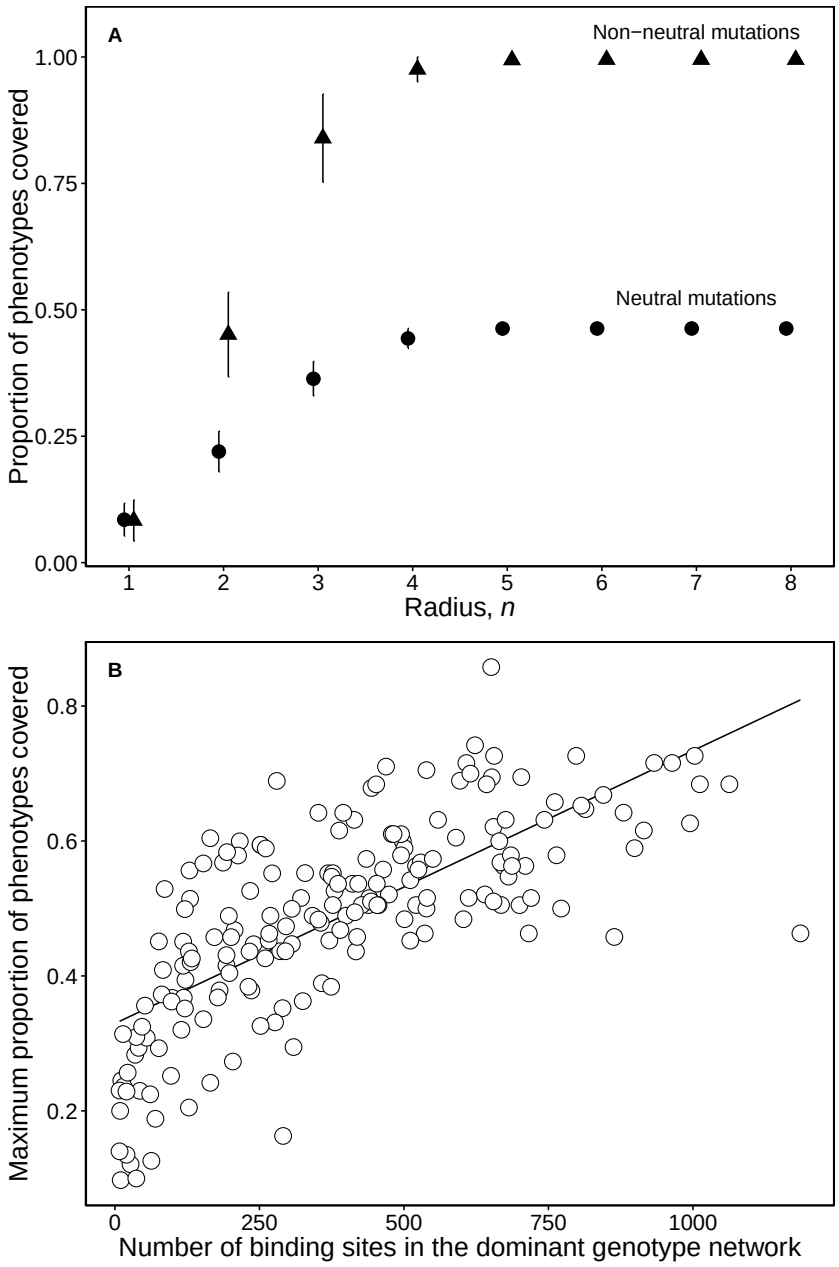


Figure 3.8: Phenotype space covering. (A) The proportion of phenotypes covered as a function of the mutational radius n from a given binding site, averaged across all binding sites of the murine TF Sp110. The maximum proportion of phenotypes covered plateaus at a much lower level when considering just neutral mutations than when considering non-neutral mutations. Error bars are the standard deviations of the mean. (B) The maximum proportion of phenotypes covered by neutral mutations as a function of the number of binding sites in the dominant genotype network, for all 190 murine TFs.

3.2.4 Genotype networks of DNA binding domains

The GP map we study can be analyzed at multiple levels of granularity. We have so far considered a fine-grained analysis, in which genotypes are DNA sequences and phenotypes are the TFs that bind these sequences. We now consider a more coarse-grained analysis, in which genotypes are DNA sequences and phenotypes are the DNA binding domains of the TFs that bind these sequences. This analysis considers a group of TFs with the same binding domain as having a single phenotype, where the genotype set of each phenotype comprises all DNA sequences that bind at least one TF with that binding domain. Studying the overlap and interface of such genotype networks complements our previous analyses by describing how TFs with different binding domains may compete for the same sites, and how DNA mutations may transfer regulatory control from a TF with one DNA binding domain to a TF with a different binding domain.

Fig. 3.9A shows the extent of overlap among all pairs of genotype networks for the 25 DNA-binding domains in the *M. musculus* dataset. Such overlap is pervasive. For example, there are six binding domains with genotype networks that overlap the genotype networks of every other binding domain in the dataset (*bHLH*, *bZIP*, *C2H2 ZFs*, *Ets*, *Homeodomain*, *SAND*). Even the *AP-2* and *Ndt80/PhoG* binding domains, which exhibit the lowest levels of overlap, still overlap with 14 (56%) of the other domains. In total, 504 of the 600 pairs of binding domains exhibit overlap in their genotype networks. It is therefore common for TFs with different binding domains to recognize some of the same sites, highlighting the potential for crosstalk in transcriptional regulation [387]. Similar patterns hold in *A. thaliana* and *N. crassa* (Supplementary Figs. 3.13A, 3.14A), even though these species have several binding domains that are not present in the *M. musculus* dataset. Such overlap therefore appears to be a general consequence of the low specificity with which eukaryotic TFs interface with DNA, rather than a consequence of the binding preferences of any particular binding domain.

Fig. 3.9B shows ϕ_{qp} for all pairs of the 25 DNA-binding domains in the *M. musculus* dataset. As with overlap, we observe an increase in ϕ_{qp} as we shift the level of analysis from

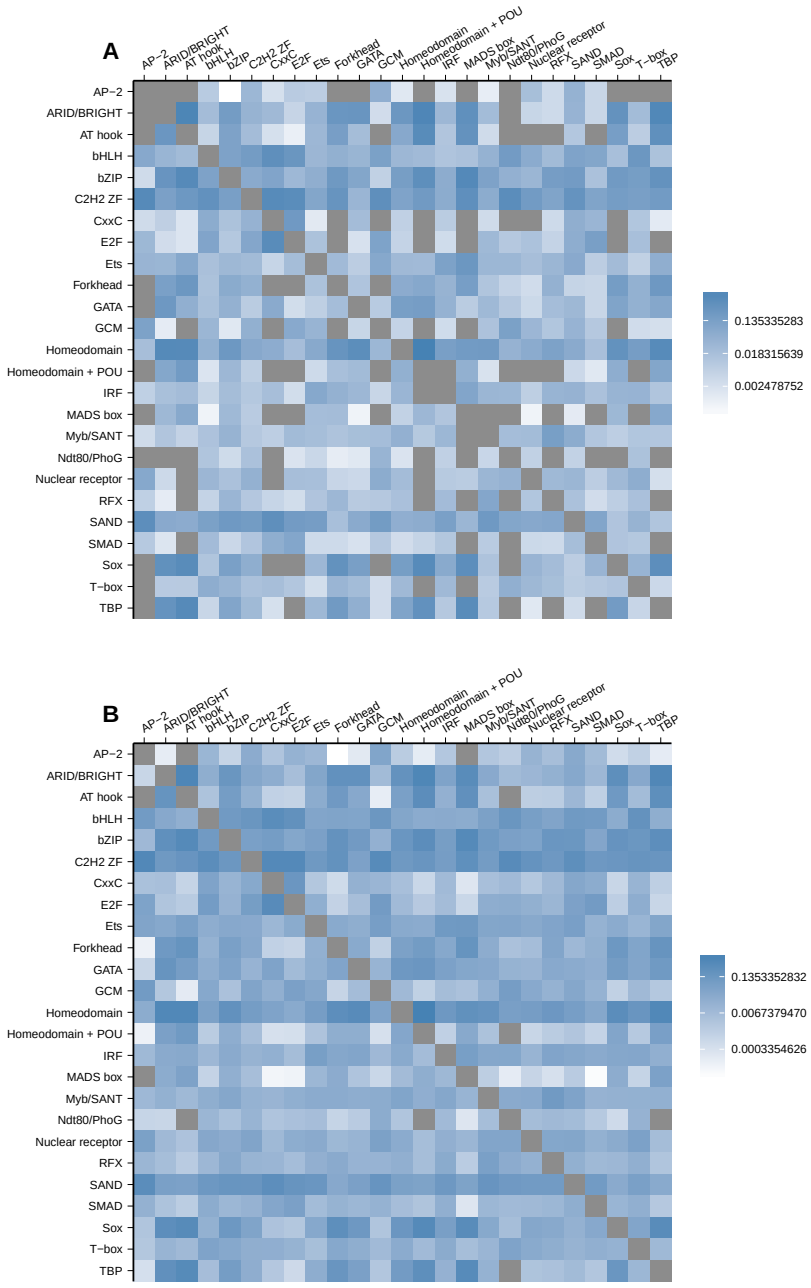


Figure 3.9: Matrices of inter-network relationships for the genotype networks of binding domains from *M. musculus*. Heatmaps of log10-transformed (A) overlap and (B) ϕ_{qp}, the probability of mutating from the genotype network of phenotype p to the genotype network of phenotype q. Each row and column represents a different genotype network. Domains are ordered alphabetically. Cells colored in gray indicate either N/A values (on the diagonal) or values equal to zero (off-diagonal).

TFs to DNA binding domains. A total of 590 (98.3%) binding domain pairs exhibit non-zero ϕ_{qp}, with values ranging from a minimum of 3.638×10^{-5} for the *Homeodomain + POU* domain and the *Homeodomain* domain, to a maximum of 0.773 for the *Forkhead* domain

and the *AP-2* domain. Mutations in TF binding sites could thus commonly transfer regulatory control among TFs with different binding domains. Similar observations are made for *A. thaliana* and *N. crassa* (Supplementary Figs. 3.13B, 3.14B). We also studied how the different genotype networks of DNA binding domains interface with one another through the visualization of phenotype networks (Figs. 3.15-3.17).

Finally, Φ_q scales with genotype network size, just as it did at the level of individual TFs (compare Supplementary Figs. 3.18 - 3.20 with Supplementary Figs. 3.10-3.12). However, since the number of TFs per binding domain in the *M. musculus* dataset also scales with genotype network size (Supplementary Fig. 3.21A), we were concerned that these trends may stem from ascertainment bias. This could occur if the number of TFs per binding domain in the *M. musculus* dataset was not representative of the number of TFs per binding domain in the *M. musculus* genome. Supplementary Figs. 3.21B,C show that this is not the case. Both the number of TFs per binding domain in the *M. musculus* dataset, and the size of the corresponding genotype network, scale with the number of TFs per binding domain in the *M. musculus* genome. We made similar observations in *A. thaliana* and *N. crassa* (Supplementary Figs. 3.22, 3.23).

3.3 Discussion

The concept of a genotype-phenotype (GP) map can be traced back to the work of Sewall Wright [40], Conrad H. Waddington [388], and John Maynard Smith [41]. However, the term GP map ("genotype-phenotype mapping") was only coined in 1970 by Jim Burns [42], who recognized the importance of incorporating a mechanistic perspective into the evolutionary framework of population genetics, thus outlining the research programme that has come to be known as evolutionary systems biology [24]. The term was re-introduced in 1991 by the developmental biologist Pere Alberch [35], who was interested in macroscopic phenotypes arising from complex developmental processes. The study of GP maps is currently shifting away from the conceptual and computational models that shaped the thinking of the founders of the field, toward empirical data derived from high-throughput assays [25–34]. Our study is part of this shift. We have used

experimental data from protein binding microarrays to analyze the architecture of an empirical GP map, in which genotypes are short DNA sequences and phenotypes are the TFs that bind these sequences. This study expands upon our previous analyses of this map [74, 156] by providing more nuanced descriptions of individual genotype networks, detailed characterizations of how these networks overlap and interface with one another, and does so at two levels of phenotypic granularity.

Our analyses of individual genotype networks provides three new insights into their structure. First, they tend to be "small-world"[382], an observation that furthers our understanding of the "robust-yet-evolvable" nature of TF binding sites [156]: While binding sites tend to be highly clustered in their genotype network (robustness), it remains possible to traverse the network with just a few mutations, thus providing efficient access to adjacent genotype networks (evolvability). Indeed, the route factor of these genotype networks indicates that they are almost optimally distributed in genotype space, in the sense that almost all genotypes are connected to a central target genotype through the shortest mutational paths. Second, these genotype networks are assortative, meaning that robust binding sites are likely to neighbor other robust binding sites. This finding has implications for the evolution of TF binding sites, because an evolving population tends to accumulate in such densely connected regions of genotype networks [50] and because assortative genotype networks can lead to "phenotypic entrapment" [68], a phenomenon in which an evolving population cannot escape its genotype network, thus frustrating evolutionary search. Third, genotype networks of TF binding sites can be partitioned in multiple meaningful ways, using either structural information or binding affinity data. Such partitions may provide valuable information about TF-DNA interactions, e.g., by helping to generate more informative TF binding motifs, or by aiding in the discovery and analysis of TFs' secondary binding preferences [195].

A commonly overlooked feature of GP maps is that genotypes may have more than one phenotype, which means that genotype networks may overlap. This is a surprising oversight, given the ubiquity of phenotypic plasticity in nature [389]. Even if we restrict

our examples to the molecular realm, they are numerous: An RNA transcript can be translated into different proteins [229], an amino acid sequence can fold into different conformational structures [390], and a promiscuous enzyme can catalyze different reactions [307]. In the GP map studied here, such overlap is also pervasive, both at the level of individual TFs and of DNA binding domains. It implies competition for binding sites among cognate and non-cognate TFs, a phenomenon known as "crosstalk." Recent modeling work suggests that crosstalk is an inevitable feature of transcriptional regulation in species that employ limited-specificity TFs [387], such as the three eukaryotic species studied here. This is important because crosstalk places constraints on the function and evolution of transcriptional regulatory networks. Our results provide an empirical complement to these earlier theoretical findings, by providing estimates of how much crosstalk can occur among TFs and binding domains. However, it is worth highlighting that these estimates are based on *in vitro* measurements of TF binding preferences. The myriad complexities of *in vivo* TF-DNA interactions [334], including epigenetic marks, local sequence and chromatin context, as well as interactions with protein partners, will certainly affect these estimates. Our ability to interrogate the effects of these complexities on TF-DNA interactions is continuing to advance [366, 391–393], and we believe that genotype networks will provide a useful framework for studying how such complexities mitigate crosstalk in transcriptional regulation.

Our analysis of how genotype networks interface with one another has implications for the emergence of evolutionary innovations, because mutations in cis-regulatory regions may produce novel gene expression patterns [166, 315]. In particular, single-base pair mutations in TF binding sites can shift the regulatory control of a gene from one TF to another, and this may cause profound phenotypic change. For example, such mutations led to the differential expression of Rhodopsin genes in different subsets of *Drosophila* photoreceptors [333], which facilitated the discrimination of a wide spectrum of optical stimuli, and thus drastically changed how flies perceive their environment. In the GP map studied here, we have previously shown that genotype networks are so intertwined

that it is usually possible to mutate at least one of a TF's binding sites to a binding site of nearly any other TF [156]. This means that mutation can readily shift the regulatory control of a gene from one TF to another, a shift that may lead to an adaptive change in gene expression. Here, we provide a more detailed and nuanced view of TF binding site evolvability. At the most local scale, evolvability is relatively low because neutral neighbors tend to have highly similar mutational neighborhoods, which decreases the diversity of novel phenotypes that may arise via a single point mutation to any one binding site [122]. However, only very few mutations are required to shift regulatory control from the cognate TF to nearly any other TF in our dataset. At even this intermediate scale, TF binding sites are therefore remarkably evolvable.

An important challenge in the biological sciences is to provide a comprehensive description of the architecture of an empirical GP map. The hyper-astronomical size of genotype space renders this challenge impossible for most biological systems of interest, including macromolecules, regulatory circuits, and metabolisms [379]. Even for the relatively small genotype space studied here, we fall short of a comprehensive description. The reason is that we do not have data describing the binding preferences of every TF from each of our three study species. However, the data we do have are a representative sampling of each species' TF repertoire. There are two reasons for this. The first is that the assayed TFs were intentionally selected to exhibit an even balance among DNA binding domains and to survey different levels of sequence similarity [196]. The second is that the number of TFs per binding domain in our dataset is correlated with the number of TFs per binding domain in the genomes of each species [394]. This study therefore provides a high-resolution depiction of the architecture of an empirical GP map.

3.4 Materials and Methods

3.4.1 Genotype networks

To construct each genotype network of TF binding sites, we followed the same procedure as Payne and Wagner [156]. First, we determined the set of sites with an E-score (a

proxy for binding affinity) higher than 0.35. Second, we used an alignment algorithm to calculate the mutational distance between all pairs of sites. Third, we used these mutational distances to define the edges of the genotype network by connecting two sites if they have a mutational distance of one.

We consider two kinds of mutations: point mutations, and indels that shift an entire, contiguous binding site by one base (Fig. 3.24). Two DNA sequences of length eight can differ by a single point mutation in $3 \times 8 = 24$ different ways, because each of the sequence's nucleotides can mutate into any one of the three other nucleotides (Fig. 3.24A,B). In addition, there are $4 \times 2 = 8$ possible indels that can separate two DNA sequences of length eight. The reason is that the indels we consider can cause a shift in either the $3'$ or $5'$ direction, and in both cases the unaligned nucleotide can comprise any one of the four possible bases (Fig. 3.24C,D). There is therefore a maximum of $24+8 = 32$ single mutations that can separate two DNA sequences of length eight.

We determine the mutational distance between two DNA sequences using the Smith-Waterman alignment algorithm, prohibiting gaps in all alignments. For two sequences s_1 and s_2, we calculate the number of mismatches $m(s_1, s_2)$ and $m(s_1, s_2^{'})$, where $s_2^{'}$ is the reverse complement of s_2. We then take the minimum of $m(s_1, s_2)$ and $m(s_1, s_2^{'})$ as the mutational distance between s_1 and s_2.

3.4.2 Intra-network measures

We used the following measures to characterize the internal structure of genotype networks [381]. We present them in the order they appear in the main text.

The *diameter* of a genotype network is the longest of the shortest mutational paths between any pair of genotypes. The *characteristic path length* is the average of the shortest paths.

The *clustering coefficient* c measures the fraction of a genotype's neighbors that are also neighbors themselves, averaged across all genotypes in a genotype network [382].

Formally, the clustering coefficient is calculated as

$$c = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2}{k_i(k_i-1)}\sum_{j,k}A_{ij}A_{ik}A_{jk}\right), \tag{3.1}$$

where n is the number of genotypes, k_i is the degree of node i, A is the adjacency matrix of the genotype network, and j and k are the neighbors of node i.

The *assortativity* r of a genotype network measures the propensity for genotypes with a similar number of neighbors to share an edge in a genotype network [383]. Assortativity ranges from -1 to 1. When $r < 0$, the network is disassortative; when $r = 0$ it is uncorrelated; and when $r > 0$, it is assortative. Assortativity is calculated as

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}, \tag{3.2}$$

where j_i and k_i are the degrees (i.e., number of connections) of the genotypes at the ends of the ith edge, and M is the number of edges in the genotype network.

The *route factor* q of a genotype network measures the average "directness" of the shortest mutational paths to a target genotype from all other genotypes in the network, relative to the shortest mutational paths to the target in Ω (the network used to describe genotype space). It is calculated as

$$q = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{l_{i,\text{target}}}{d_{i,\text{target}}}, \tag{3.3}$$

where n is the number of nodes in the network, $l_{i,\text{target}}$ is the shortest mutational path between genotype i and the target genotype in the genotype network, and $d_{i,\text{target}}$ is the shortest mutational path between genotype i and the target genotype in Ω [384]. We use the highest-affinity binding site as the target genotype.

3.4.2.1 The stochastic block model for network partitioning

The stochastic block model (SBM) is a probabilisitic generative model for networks [395, 396]. Under the SBM, all nodes are assigned to one of k groups, and the probability of an edge connecting any pair of nodes depends only upon the nodes' group memberships. The pattern of edges can therefore be described by a single $k \times k$ "mixing matrix," in which each element p_{rs} gives the interaction probability between groups r and s (i.e., the probability that an edge exists between a node from group r and a node from group s.)

Using statistical inference [397], we determined the maximum likelihood group assignment for each of the nodes in each genotype network. For a given assignment, the maximum likelihood interaction probability between groups r and s is given by the observed number of edges between the groups divided by the number of possible edges between the groups. That is,

$$p_{rs} = \frac{e_{rs}}{n_r n_s}, \tag{3.4}$$

where e_{rs} is the number of edges connecting nodes in group r to nodes in group s, and n_r and n_s are the number of nodes in groups r and s, respectively. Because we set the number of groups to $k = 2$, we have just three group interaction probabilities p_{11}, p_{12} and p_{22}, because the network is undirected. By comparing these probabilities, we can determine the type of structure the groups represent. For the two-group case there are three possibilities: $p_{11} > p_{12} < p_{22}$ (assortative), $p_{11} < p_{12} > p_{22}$ (disassortative), and $p_{11} > p_{12} > p_{22}$ (core-periphery) [385].

Introduced in [386], the block model entropy significance test provides a means for identifying whether node attributes are distributed randomly across a network. The test works by partitioning a network into groups of nodes that have the same node attribute value (for continuous-valued attributes, we form groups by discretizing the values into bins). Using this partition $\mathcal{M}$, we calculate the mixing matrix using Eq. (3.4). As a test statistic, we calculate the SBM entropy $\mathcal{H}$,

$$\mathcal{H}(\mathcal{M}) = -\left[\sum_{rs} e_{rs} \log p_{rs} + (n_r n_s - e_{rs} \log(1 - p_{rs})\right]. \tag{3.5}$$

High entropy indicates that node attributes are not correlated with network structure. Low entropy indicates that there is a correlation between the node attributes and the network structure. To determine if this correlation is statistically significant, we compare the observed entropy against a null distribution of entropy values. We obtain this distribution by randomly permuting node attributes, resulting in new partitions $\{\pi\}$ and corresponding mixing matrices. Importantly, this choice of null model preserves both the observed network structure and the relative frequencies of attribute values, but removes any correlation between the two. The result is a standard empirical p-value, defined as

$$p = Pr[H(\pi) \leq H(\mathcal{M})]. \tag{3.6}$$

Smaller p-values indicate a lower plausibility that a random permutation of the node attributes could describe the network structure as well as the observed distribution of node attributes.

3.4.2.2 Binding affinity partitions

We used the SBM partitions as a baseline for building node partitions that are based on binding affinities. For each genotype network, we attributed a categorical label to every node, indicating its SBM group. We chose "0" for nodes in the most assortative group and "1" for other nodes. This labeling also induces a partitioning of the binding affinities into two groups.

For each genotype network, we performed a logistic regression of the SBM partitioning of binding affinities. Using $b_{\min}$ and $b_{\max}$ to denote the minimum and maximum binding affinity values within a given genotype network, the regression resulted in a classifier $C : [b_{\min}, b_{\max}] \to [0, 1]$ that we trained on the empirical data. This classifier provided the likelihood that a given binding affinity value belonged to one SBM group or the other. In order to distinguish between "high" and "low" binding affinities, we chose the critical value b^* to be the binding affinity at which the classifier distinguished between groups, i.e. $C(b^*) = 0.5$. We used b^* in order to obtain a binding affinity partition, with nodes

having binding affinities less than or equal to b^* in a group labeled g_{low}, and nodes with binding affinities greater than or equal to b^* in another group labeled g_{high}.

To test the statistical significance of binding affinity with respect to the structure of a genotype network, we again used the block model entropy significance test, such that in Eqs. 3.4 and 3.5, groups r and s were replaced with groups g_{low} and g_{high}.

3.4.3 Inter-network measures

We characterized the arrangement of genotype networks in genotype space by measuring overlap and mutation probabilities ϕ_{qp} among all pairs of phenotypes. We applied these measures at two levels of phenotypic granularity. In the first, the phenotype of a binding site genotype is the TF that binds the site. In the second, the phenotype of a binding site genotype is the DNA binding domain that binds the site. Regardless of the definition of phenotype, the measures were applied to the corresponding genotype networks in the same way.

The *overlap* $\mathcal{O}_{qp}$ of dominant genotype networks G_p and G_q, corresponding to phenotypes p and q, is defined as

$$\mathcal{O}_{qp} = \frac{|S(G_q) \cap S(G_p)|}{|S(G_p)|}, \tag{3.7}$$

where $S(G_q)$ is the set of genotypes in genotype network G_q, and $|S(G_q)|$ is the number of genotypes in this set. Note that overlap is an *asymmetric* measure due to the normalization factor corresponding to the number of binding sites in G_p.

The fraction ϕ_{qp} of mutations to binding sites in genotype network G_p that create binding sites in genotype network G_q is defined as

$$\phi_{qp} = \frac{1}{|S(G_p)|} \sum_{i \in S(G_p)} \phi_q^{\text{local}}(i), \tag{3.8}$$

where

$$\phi_q^{\text{local}}(i) = \frac{n_i^q}{k_i}, \tag{3.9}$$

n_i^q is the number of neighbors of genotype i that have phenotype q, and k_i is the number of neighbors of genotype i in Ω. Thus, $\phi_q^{\text{local}}(i)$ is the fraction of genotype i's neighbors that have phenotype q. We use Eqs. 3.8 and 3.9 to calculate the mutational connectivity Φ_q of the genotype network of phenotype q from the genotype networks of all other phenotypes in genotype space as

$$\Phi_q = \sum_p \phi_{qp}. \tag{3.10}$$

The measure ϕ_{qp} is similar to the phenotypic accessibility A_{qp} of phenotype q from phenotype p, which is measured as

$$A_{qp} = \frac{|S(G_q) \cap \partial S(G_p)|}{|\partial S(G_p)|}, \tag{3.11}$$

where $S(G_q)$ is the set of genotypes in the dominant network of phenotype q and $\partial S(G_p)$ is the set of 1-mutant neighbors of the set $S(G_p)$ [70, 152]. We compute this measure simply as a point of comparison with ϕ_{qp}.

We complemented these global inter-network comparisons by comparing the phenotypic compositions of the local mutational neighborhoods of genotype pairs (i, j), using the Bhattacharyya coefficient [122]:

$$BC(i, j) = \sum_q \sqrt{\phi_q^{\text{local}}(i) \times \phi_q^{\text{local}}(j)}. \tag{3.12}$$

This coefficient ranges from a minimum of zero when the phenotypic compositions of the mutational neighborhoods of genotypes i and j are maximally dissimilar to a maximum of one when they are identical. To quantify whether the phenotypic compositions of mutational neighborhoods are more similar among pairs of genotypes (i, j) that are neutral neighbors than among pairs of genotypes (i, k) that are not neutral neighbors, but are from the same genotype network, we computed the *similarity ratio* of the Bhattacharyya coefficients $BC(i, j)/BC(i, k)$. A ratio greater than 1 indicates that the phenotypic compositions of mutational neighborhoods of pairs of genotypes are more similar if those genotypes are connected by a neutral mutation than if they are not, and vice versa.

Neighbors that are shared amongst genotypes i and j, and amongst i and k, are excluded from this analysis to provide a more conservative measure.

3.4.4 Determining the number of TFs per DNA binding domain

We obtained the proteomes of *A. thaliana* (UP000006548), *N. crassa* (UP000001805), and *M. musculus* (UP000000589) from UniProt [394]. To find the number of proteins in each proteome with a match to a DNA-binding domain, we employed the program hmmsearch from the software package HMMER (v3.1b2) (*http://hmmer.org/*). We used a cutoff of 0.01 for both the sequence e-value and the domain conditional e-value. We downloaded the hidden Markov models of each DNA-binding domain from the Pfam database (v27.0) [398].

3.5 Supplementary results

3.5.1 Some sequences have fewer than 32 neighbors in genotype space

Of the 32,896 sequences in genotype space, 1,312 have fewer than 32 neighbors (Fig. 3.1D). This occurs when two different mutations to a sequence yield the same mutated sequence, forcing the prioritization of one mutation over another. If the mutations are of different types (i.e., a point mutation and an indel), we always prioritize the point mutation over indels because laboratory evolution experiments indicate that they occur more frequently than indels [399, 400].

The 1,312 genotypes with fewer than 32 neighbors fall into the following five groups:

1. There are $4^4 = 256$ sequences that are identical to their reverse complements, and 252 of these have 16 neighbors. Due to the symmetry of these sequences, the number of possible point mutations in them is reduced from 24 to 12. The reason is that a point mutation in position $i < 4$ is equivalent to a point mutation to the Watson-Crick pair in position $7 - i$, after taking the reverse complement. For example,

consider the point mutation `A` $\rightarrow$ `C` in the 0th position of `ACGTACGT`. This yields the same sequence (`CCGTACGT`) as a point mutation `T` $\rightarrow$ `G` in the 7th position, after taking the reverse complement of the mutated sequence. The symmetry of these sequences also reduces the number of possible indels from 8 to 4. For example, an indel separates the sequence `ACGTACGT` from `CGTACGTA`, such that an alignment will leave the 0th position of the former sequence and the 7th position of the latter sequence unaligned. An indel also separates the sequence `ACGTACGT` from `TACGTACG`, such that an alignment will leave the 7th position of the former sequence and the 0th position of the latter sequence unaligned. Since the sequences `CGTACGTA` and `TACGTACG` are reverse complements of one another, it is not possible for `ACGTACGT` to have both of these mutational neighbors. In sum, these 252 sequences only have $12 + 4 = 16$ neighbors.

2. Of the 256 sequences that are identical to their reverse complements, four have 15 neighbors: `AAAATTTT`, `CCCCGGGG`, `GGGGCCCC`, `TTTTAAAA`. The reasons are the same as for the other 252 sequences, except that the number of possible indels is further reduced to 3. To understand why, consider aligning the sequence `AAAATTTT` with `AAATTTTT`. This alignment could either include a point mutation in the 3rd position, or an indel that leaves the 0th position of the former sequence and 7th position of the latter sequence unaligned. For this reason, these four sequences have 15 neighbors.

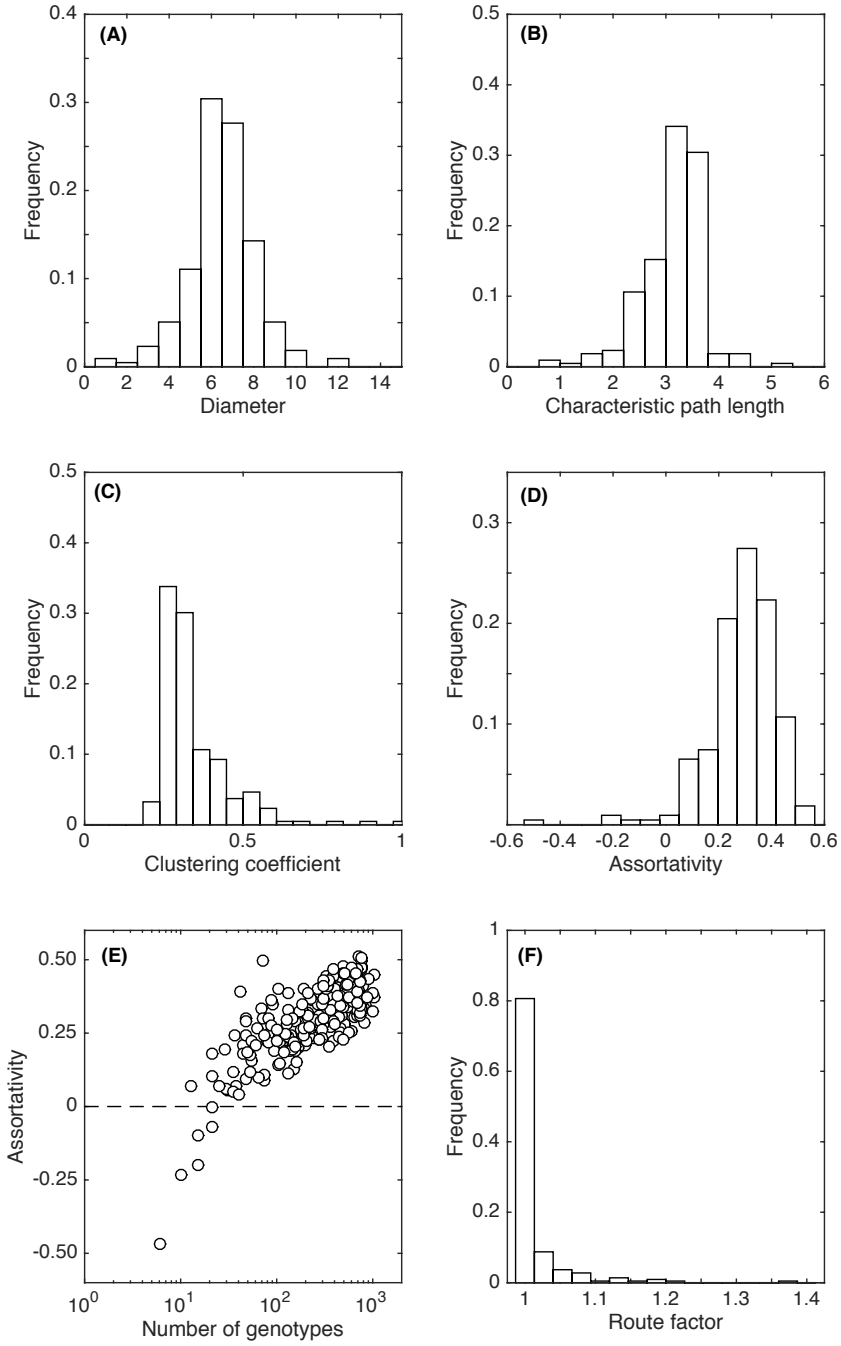
3. There are two sequences with 24 neighbors: `AAAAAAAA` and `CCCCCCCC`. They have 24 neighbors because we prioritize point mutations: We consider that any mutation that might be caused by an indel is more likely to be caused by a point mutation.

4. There are 46 sequences with 30 neighbors. 41 of these are of the form `AAAAAAAC`, `AAAAAACC`, `AAAAACCC`, ... , `ACCCCCCC`, for which the number of possible indels is reduced from 8 to 6 because 2 indels are superseded by point mutations. For example, consider the sequence `AAAAAAAC`, which can be aligned to the sequences `AAAAAACC` and `AAAAAAAA` using either a point mutation or an indel. The remaining five se-

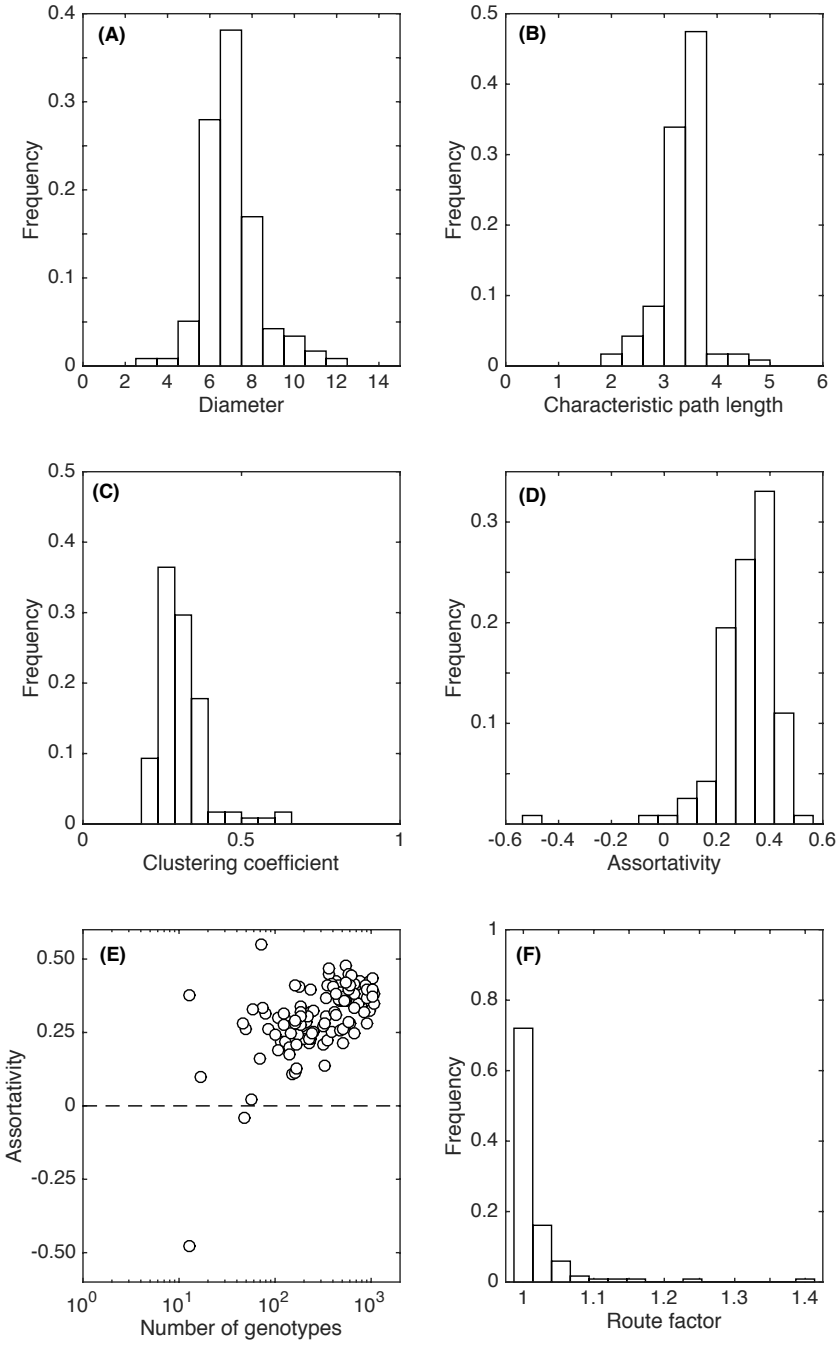
quences are ACACTGTG, AGAGTCTC, ATATTATA, CTCTGAGA, GTGTCACA. These sequences also have the number of indels reduced from 8 to 6, but for a more complicated reason. As an example, consider the sequence ACACTGTG, which is separated by a single point mutation from ACACAGTG. The reverse complement of ACACAGTG is CACTGTGT, which can be aligned to ACACTGTG with an indel, a mutation that is superseded by the point mutation from ACACTGTG to ACACAGTG.

5. There are 1008 sequences with 31 neighbors. These sequences have one indel that is superseded by a point mutation. For example, consider the sequence AAAACTTT. A point mutation C $\rightarrow$ G in the 4th position results in the sequence AAAAGTTT, whose reverse complement AAACTTTT can be aligned to AAAACTTT via an indel. This indel is therefore not included in the neighborhood of AAAACTTT, reducing the number of neighbors to 31.

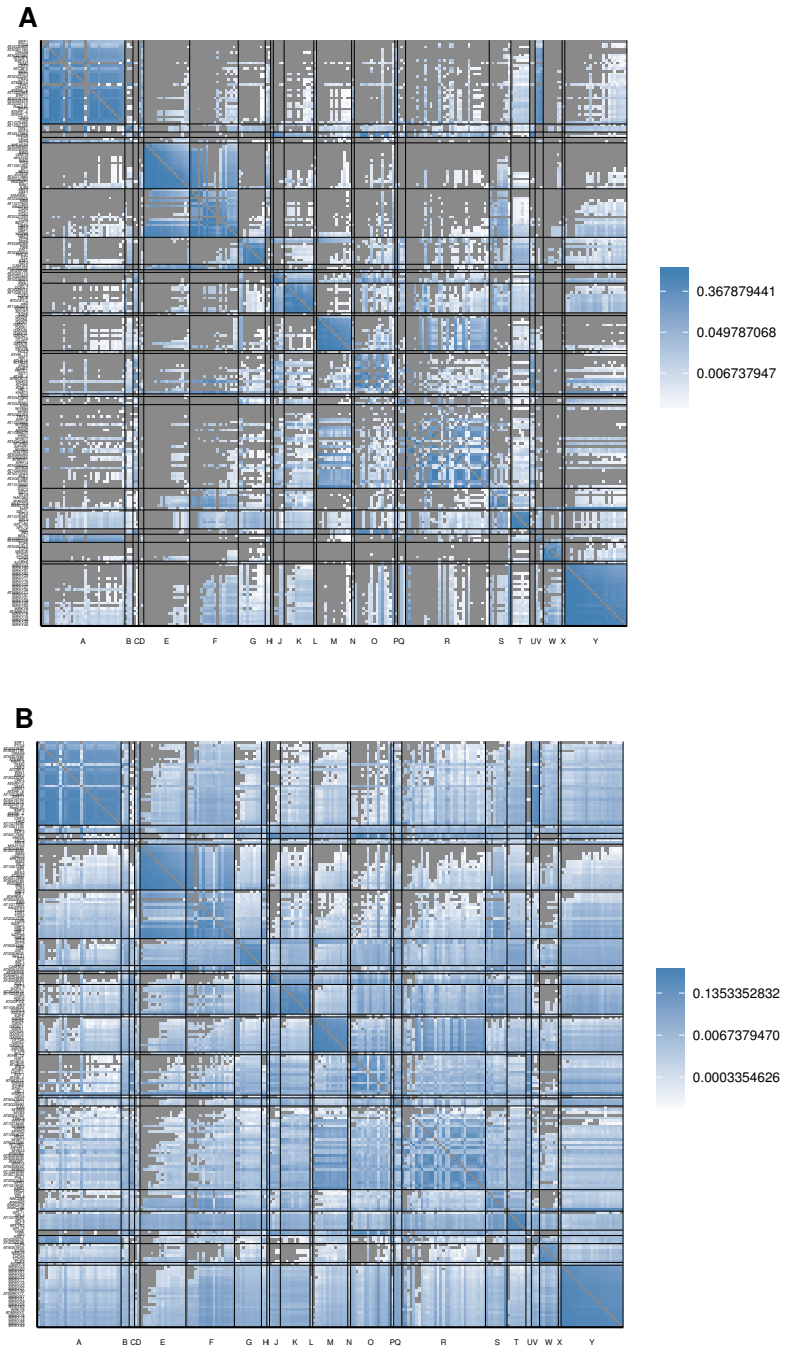
3.6 Supplementary figures



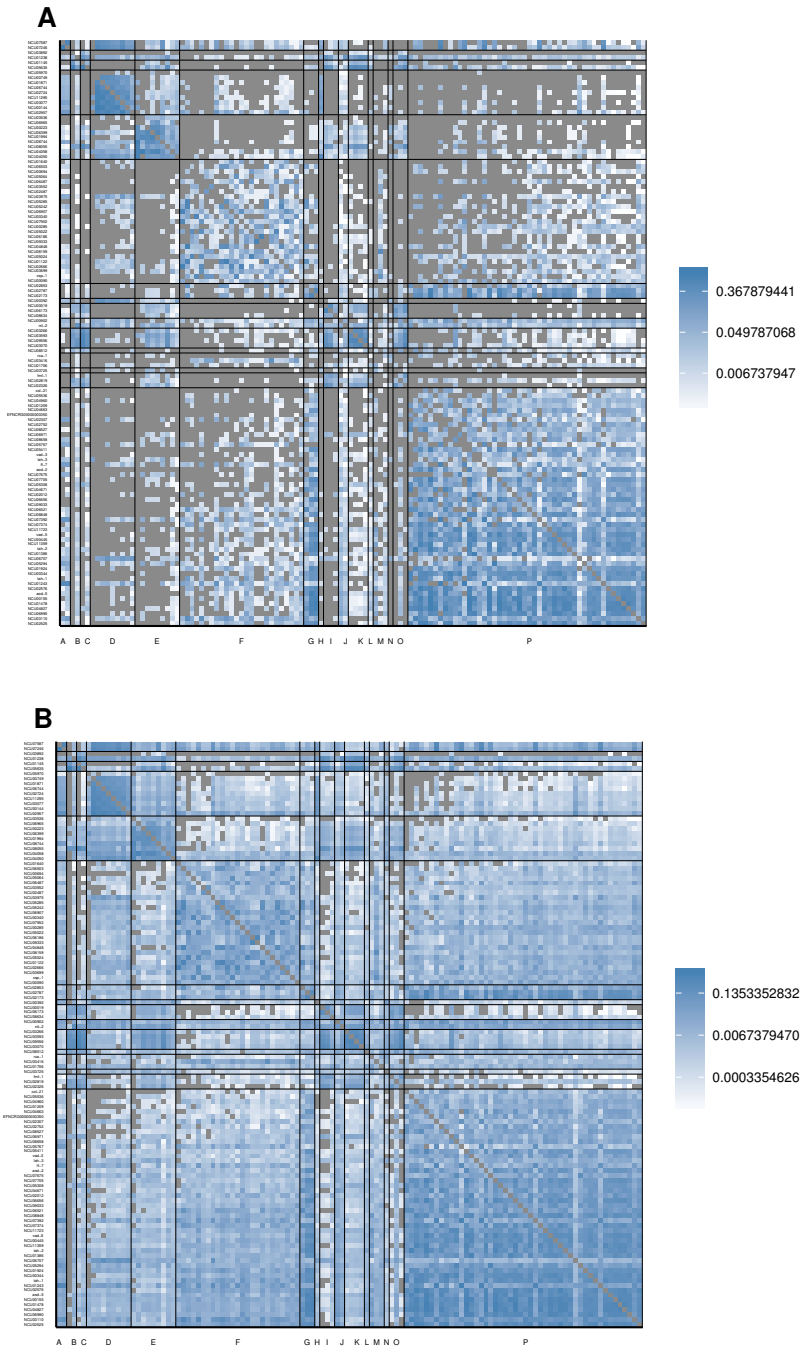
Supplementary Figure 3.1: Intra-network statistics for 218 TFs from *A. thaliana*. The distributions of genotype network (A) diameter, (B) characteristic path length, (C) clustering coefficient, and (D) assortativity. (E) Assortativity (horizontal axis) and its relationship to the number of genotypes in the dominant genotype network (vertical axis). The horizontal dashed line indicates an uncorrelated (non-assortative) mixing pattern. (F) The distribution of the genotype network route factor.



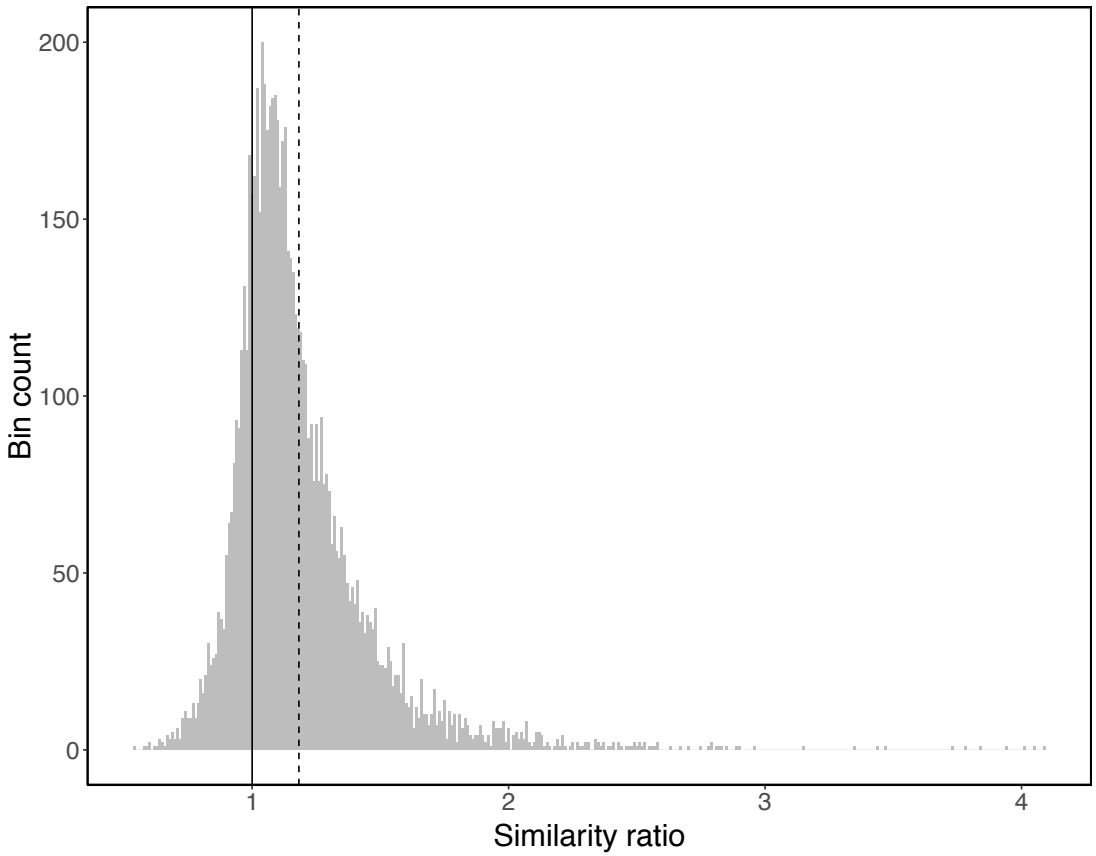
Supplementary Figure 3.2: Intra-network statistics for 119 TFs from *N. crassa*. The distributions of genotype network (A) diameter, (B) characteristic path length, (C) clustering coefficient, and (D) assortativity. (E) Assortativity (horizontal axis) and its relationship to the number of genotypes in the dominant genotype network (vertical axis). The horizontal dashed line indicates an uncorrelated (non-assortative) mixing pattern. (F) The distribution of the genotype network route factor.



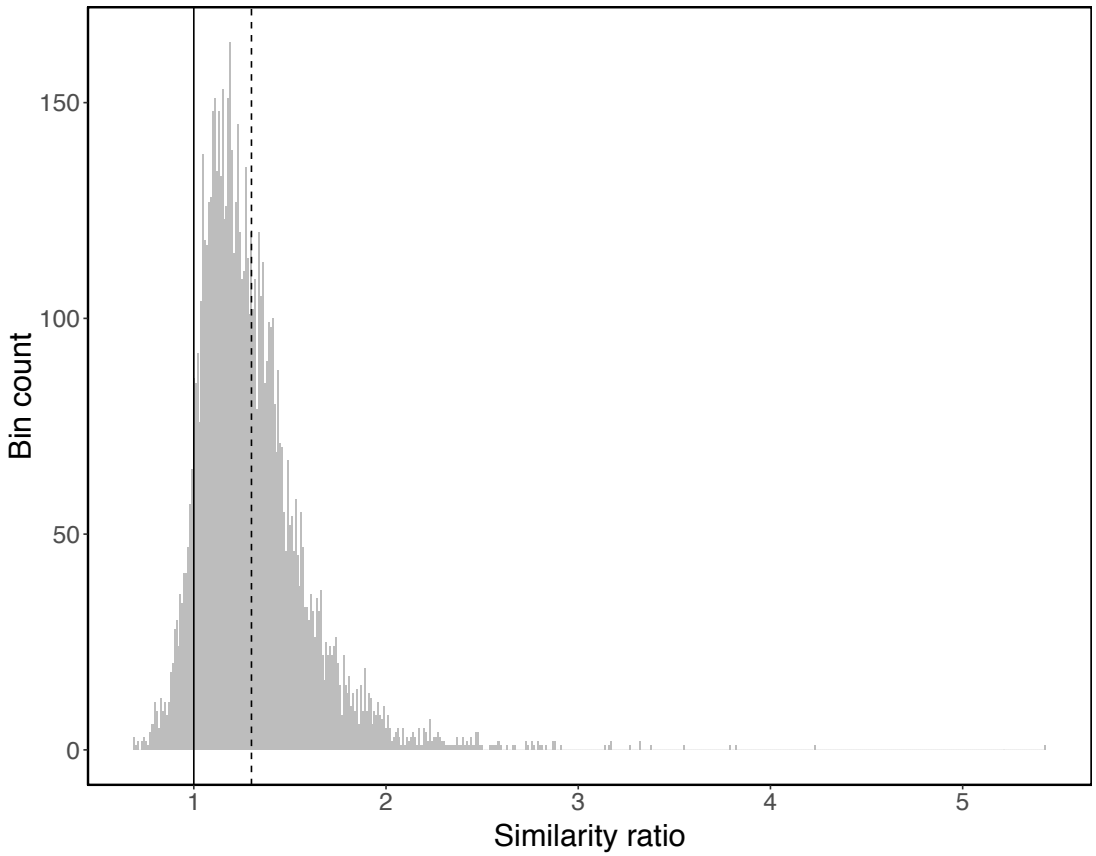
Supplementary Figure 3.3: Matrices of inter-network relationships for the genotype networks of TF binding sites from *A. thaliana*. Heatmaps of log10-transformed (A) overlap and (B) ϕ_{qp}, the probability of mutating from the genotype network of phenotype p to the genotype network of phenotype q. The rows and columns are grouped according to binding domain, which are ordered alphabetically on the horizontal axis: A, AP2; B, AP2B3; C, AT hook; D, B3; E, bHLH; F, bZIP; G, C2H2 ZF; H, CG-1; I, CSD; J, CxC; K, Dof; L, E2F; M, GATA; N: GRAS; O, Homeodomain; P, LOB; Q, MADF; R, Myb/SANT; S, NAC/NAM; T, SBP; U, Sox; V, Storekeeper; W, TCP; X, WRC; Y, WRKY. Within the DNA-binding domain groups, the rows and columns are ordered by the size of each TF's dominant genotype network, such that network size increases from top to bottom and from left to right. Labels on the vertical axis indicate the name of the TFs. Cells colored in gray indicate either N/A values (on the diagonal) or values equal to zero (off-diagonal).



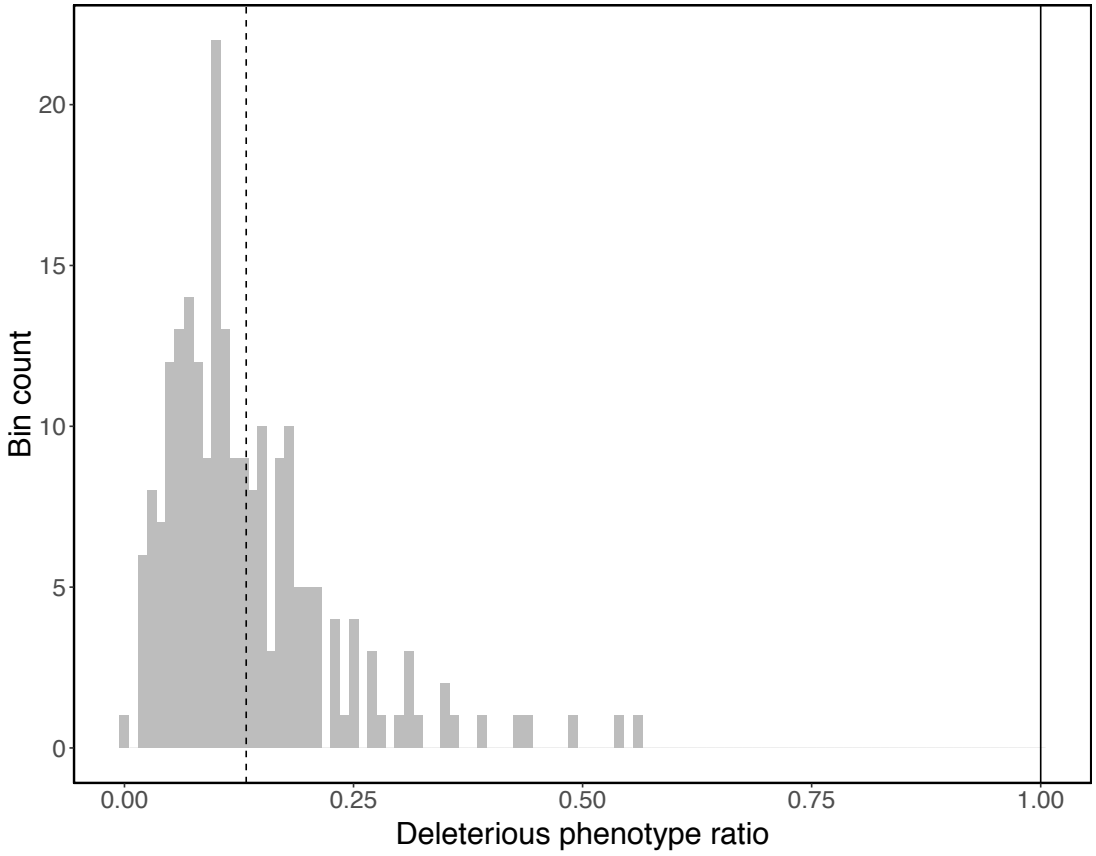
Supplementary Figure 3.4: Matrices of inter-network relationships for the genotype networks of TF binding sites from *N. crassa*. Heatmaps of log10-transformed (A) overlap and (B) ϕ_{qp}, the probability of mutating from the genotype network of phenotype p to the genotype network of phenotype q. The rows and columns are grouped according to binding domain, which are ordered alphabetically on the horizontal axis: A, APSES; B, ARID/BRIGHT; C, AT hook; D, bHLH; E, bZIP; F, C2H2 ZF; G, C2H2 ZF + Zinc cluster; H, CENPB; I, Forkhead; J, GATA; K, Homeodomain; L, HSF; M, Myb/SANT; N, Ndt80/PhoG; O, Sox; P, Zinc cluster. Within the DNA-binding domain groups, the rows and columns are ordered by the size of the dominant genotype network, such that network size increases from top to bottom and from left to right. Labels on the vertical axis indicate the name of the TFs. Cells colored in gray indicate either N/A values (on the diagonal) or values equal to zero (off-diagonal).



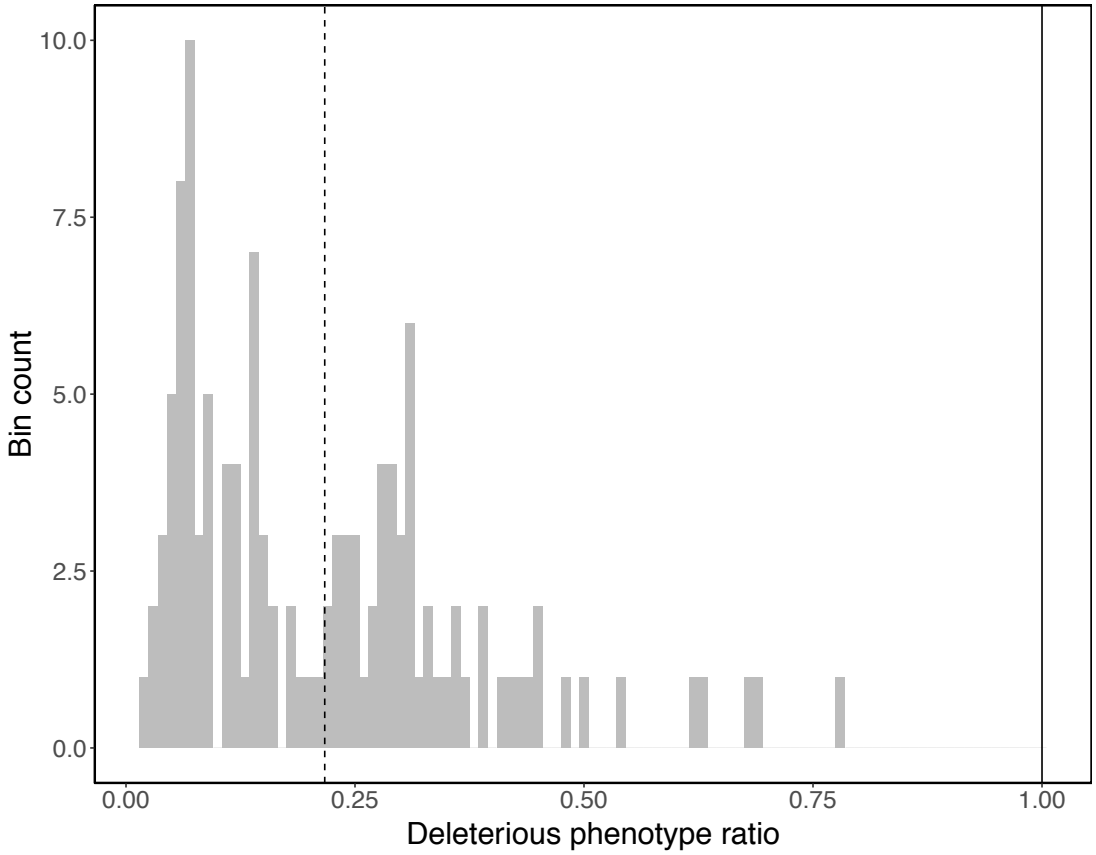
Supplementary Figure 3.5: In *A. thaliana*, **the phenotypes in the mutational neighborhoods of neutral neighbors are more similar than those of neutral pairs that are not neighbors**. The distribution of the similarity ratio (Eq. 3.12) of the phenotype probability distributions (Eq. 3.9) is shown for neutral neighbors (n_1 and n_2) and for neutral pairs that are not neighbors (n_1 and n_3). For this analysis, we considered all 7,098 pairs of neutral neighbors in the genotype network for AZF2, and sampled the same number of neutral pairs that are not neighbors. The mutational neighborhoods of n_1 and n_2 are more similar than those of n_1 and n_3, because the mean of the ratio (vertical dashed line) is larger than unity (vertical solid line). The standard error of this mean (0.003) is minute compared to the difference between the mean and unity (0.182).



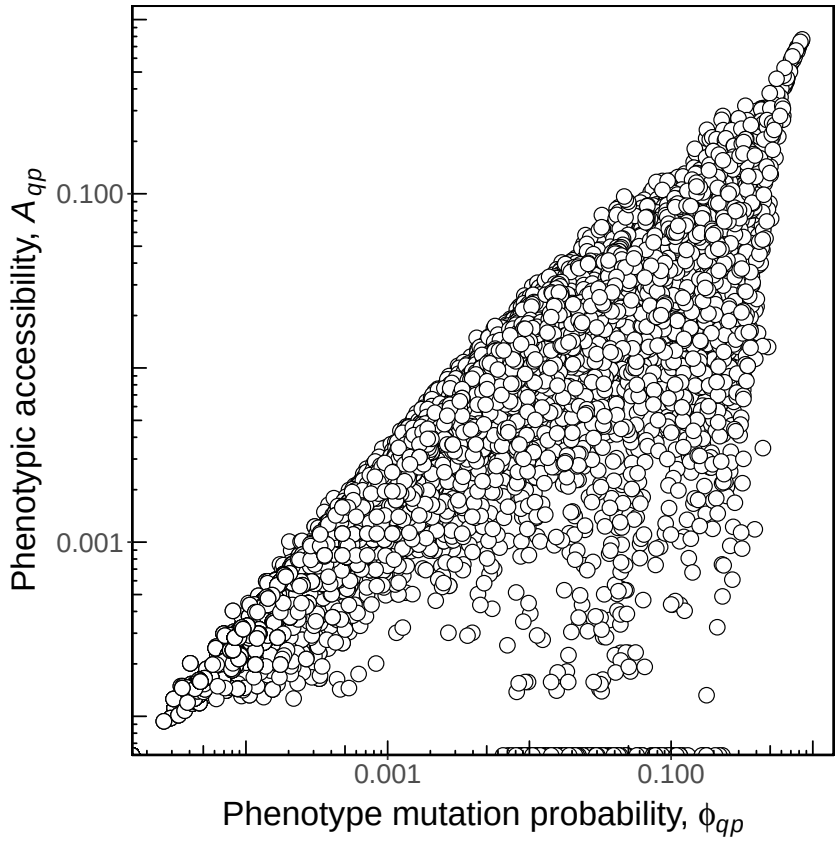
Supplementary Figure 3.6: In *N. crassa*, the phenotypes in the mutational neighborhoods of neutral neighbors are more similar than those of neutral pairs that are not neighbors. The distribution of the similarity ratio (Eq. 3.12) of the phenotype probability distributions (Eq. 3.9) is shown for neutral neighbors (n_1 and n_2) and for neutral pairs that are not neighbors (n_1 and n_3). For this analysis, we considered all 7,379 pairs of neutral neighbors in the genotype network for NCU02525, and sampled the same number of neutral pairs that are not neighbors. The mutational neighborhoods of n_1 and n_2 are more similar than those of n_1 and n_3, because the mean of the ratio (vertical dashed line) is larger than unity (vertical solid line). The standard error of this mean (0.003) is minute compared to the difference between the mean and unity (0.3).



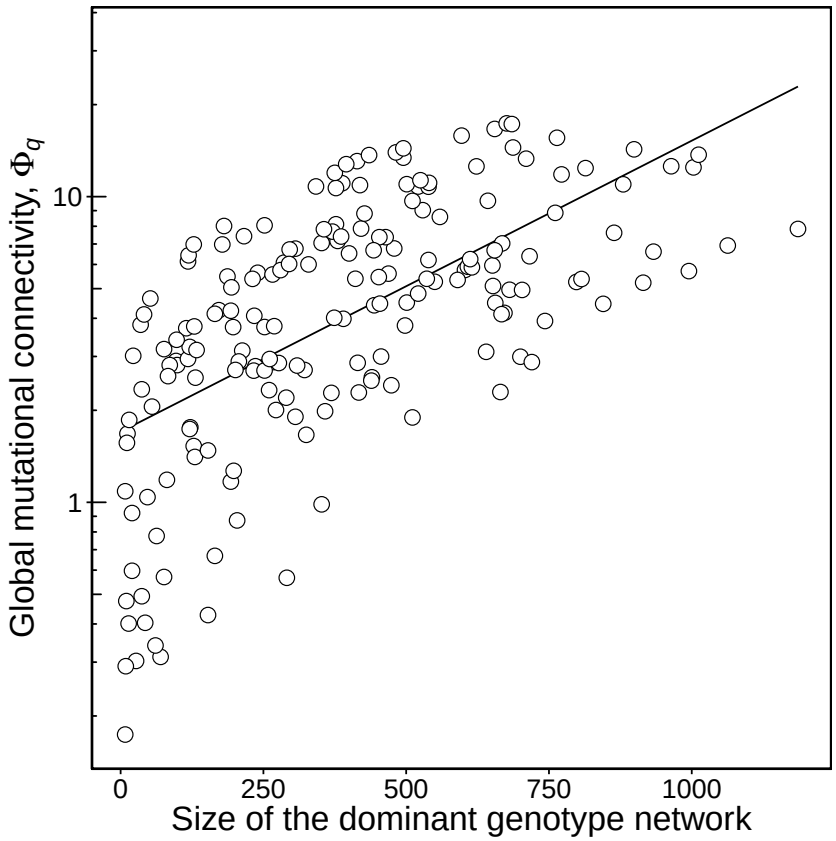
Supplementary Figure 3.7: In *A. thaliana*, unbound sites are underrepresented in the neighborhoods of bound sites. The distribution of the ratio $\phi_{\text{unbound},p}/f_{\text{unbound}}$, which is the probability of mutating from a sequence bound by TF p to an unbound sequence, divided by the null expectation of the frequency of unbound DNA sequences. The distribution is clearly skewed to values smaller than one, as shown by the distance of the distribution (vertical dashed line) to unity (vertical solid line).



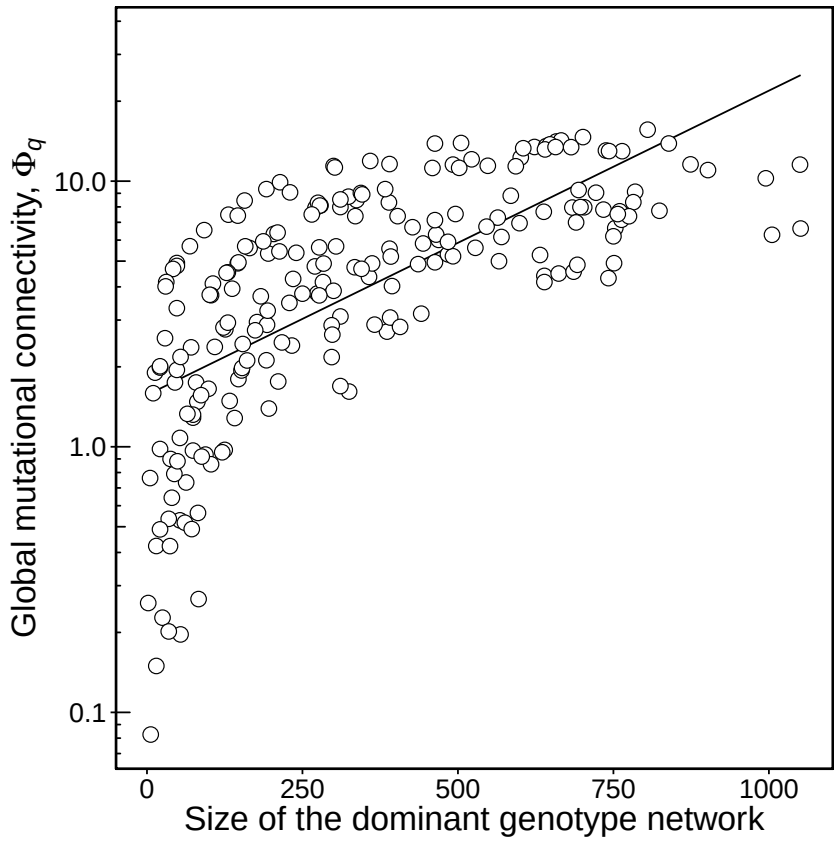
Supplementary Figure 3.8: In *N. crassa*, unbound sites are underrepresented in the neighborhoods of bound sites. The distribution of the ratio $\phi_{\text{unbound},p}/f_{\text{unbound}}$, which is the probability of mutating from a sequence bound by TF p to an unbound sequence, divided by the null expectation of the frequency of unbound DNA sequences. The distribution is clearly skewed to values smaller than one, as shown by the distance of the distribution (vertical dashed line) to unity (vertical solid line).



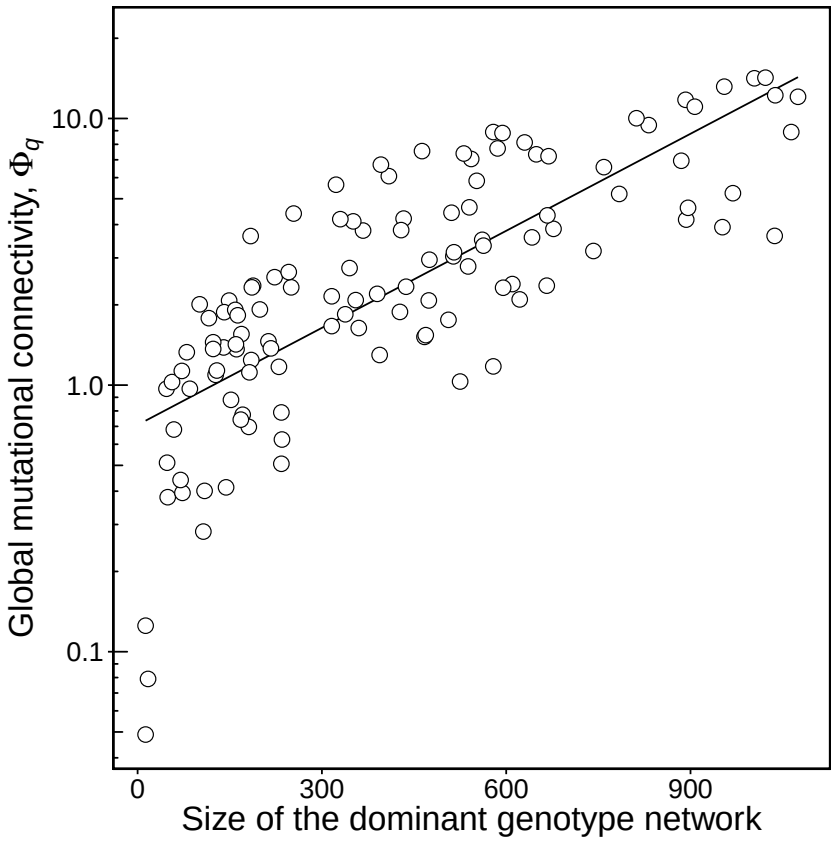
Supplementary Figure 3.9: Phenotypic accessibility A_{qp} is strongly correlated with ϕ_{qp} (Spearman's $r = 0.95$, $p < 10^{-6}$). Each circle represents one of the 35,910 pairs of TFs from *M. musculus*. Half circles at the bottom of the panel denote pairs of phenotypes with phenotypic accessibility = 0. Note the logarithmic scale on both axes.



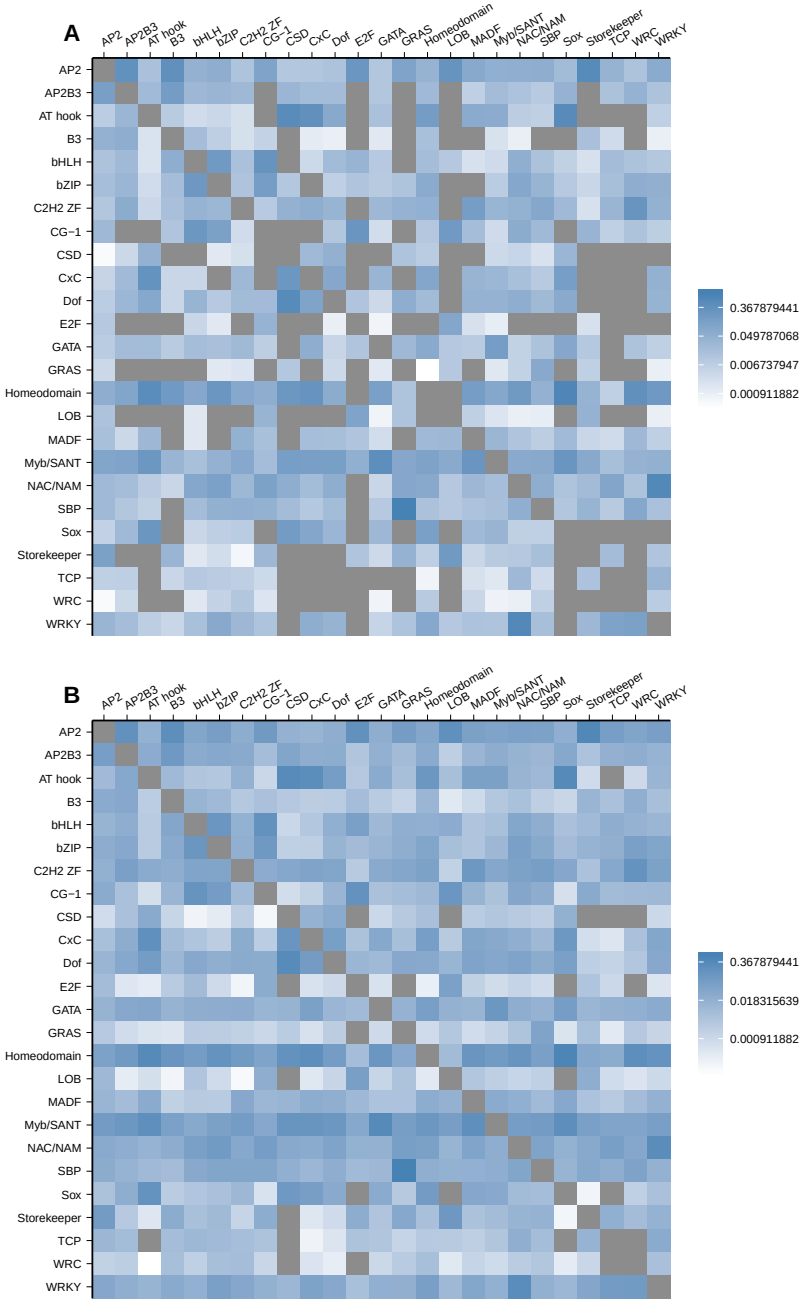
Supplementary Figure 3.10: In *M. musculus*, the global mutational connectivity of a phenotype increases with the size of its dominant genotype network. Each circle shows the global mutational connectivity Φ_q of one of the 190 *M. musculus* TFs, as a function of the number of binding sites in its dominant genotype network. The solid line is the best linear fit to the data and is provided as a visual guide.



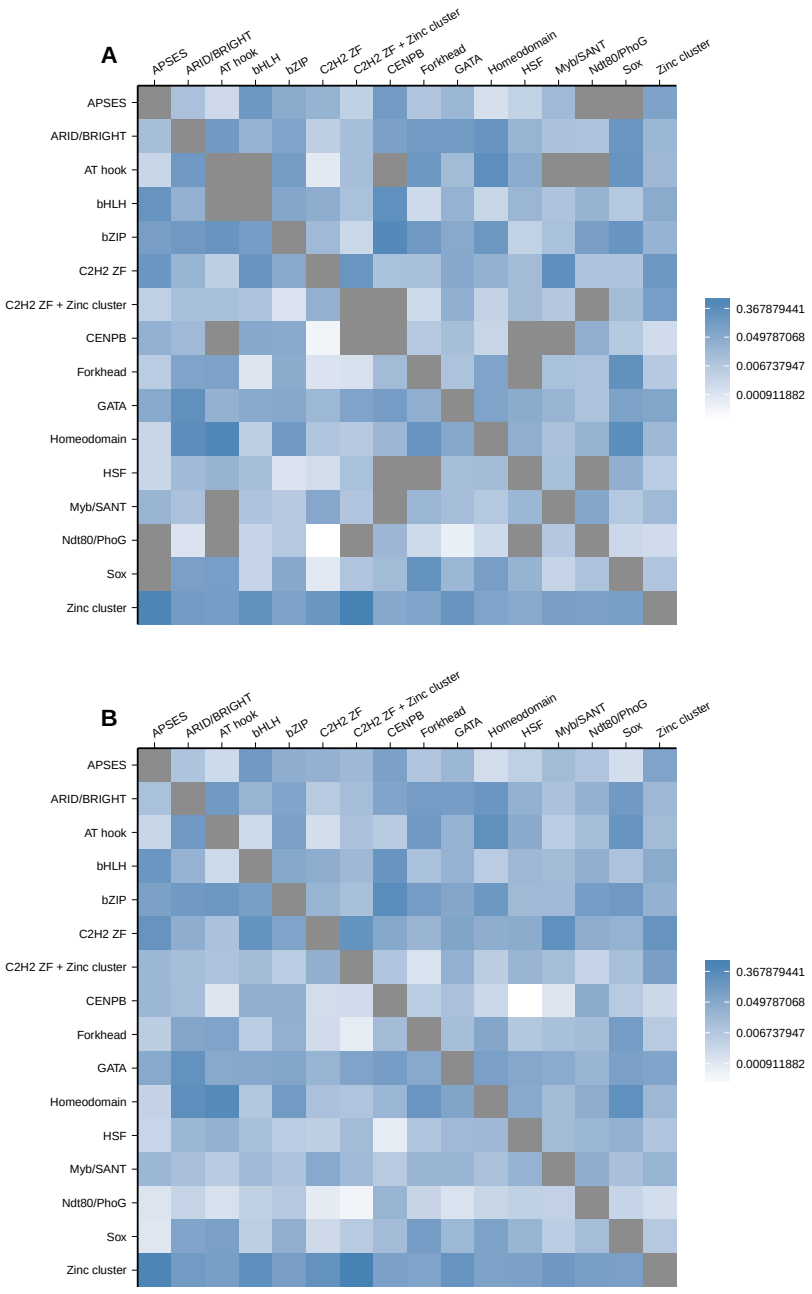
Supplementary Figure 3.11: In *A. thaliana*, the global mutational connectivity of a phenotype increases with the size of its dominant genotype network. Each circle shows the global mutational connectivity Φ_q of one of the 217 *A. thaliana* TFs, as a function of the number of binding sites in its dominant genotype network. The solid line is the best linear fit to the data and is provided as a visual guide.



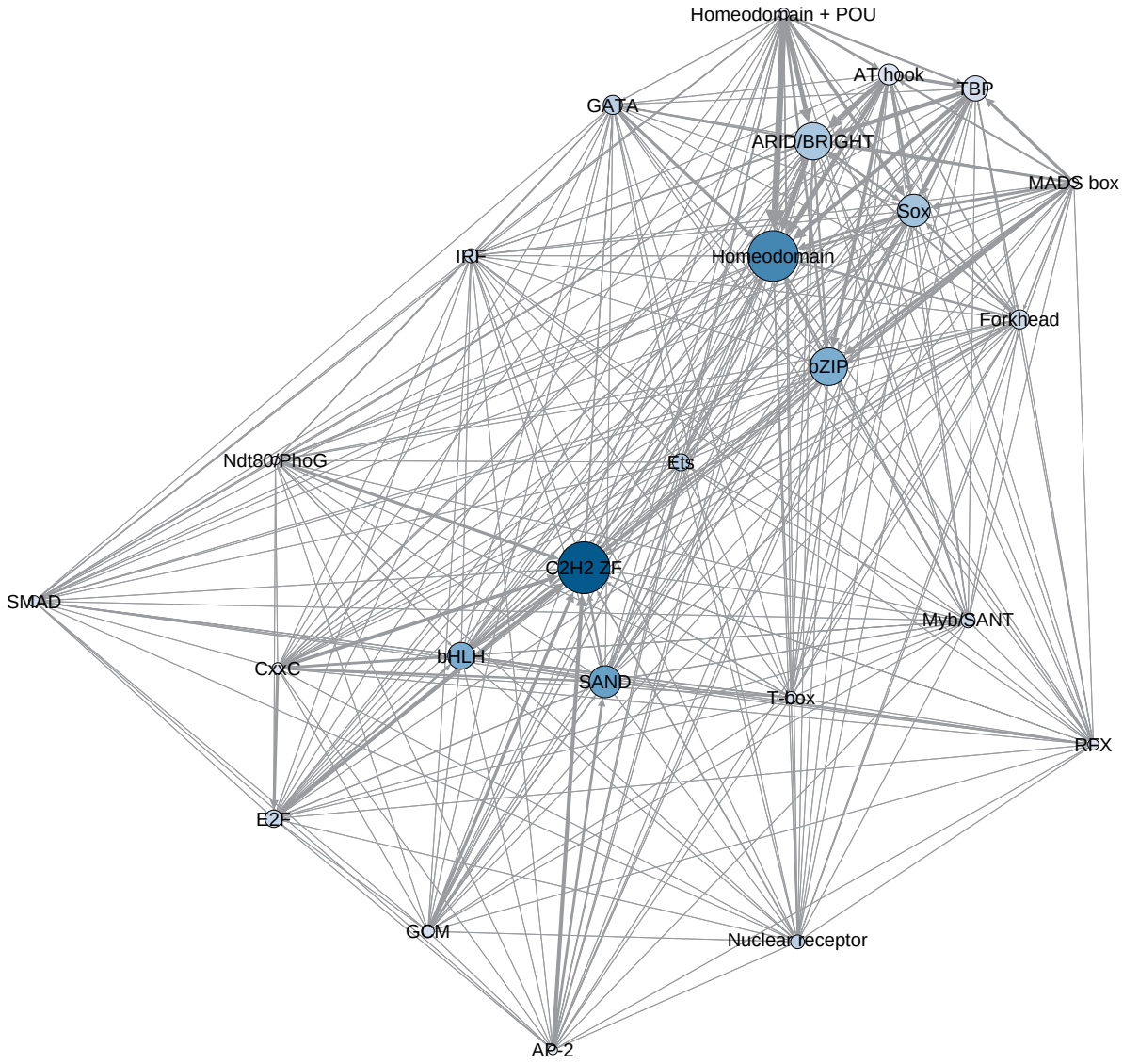
Supplementary Figure 3.12: In *N. crassa*, the global mutational connectivity of a phenotype increases with the size of its dominant genotype network. Each circle shows the global mutational connectivity Φ_q of one of the 118 *N. crassa* TFs, as a function of the number of binding sites in its dominant genotype network. The solid line is the best linear fit to the data and is provided as a visual aid.



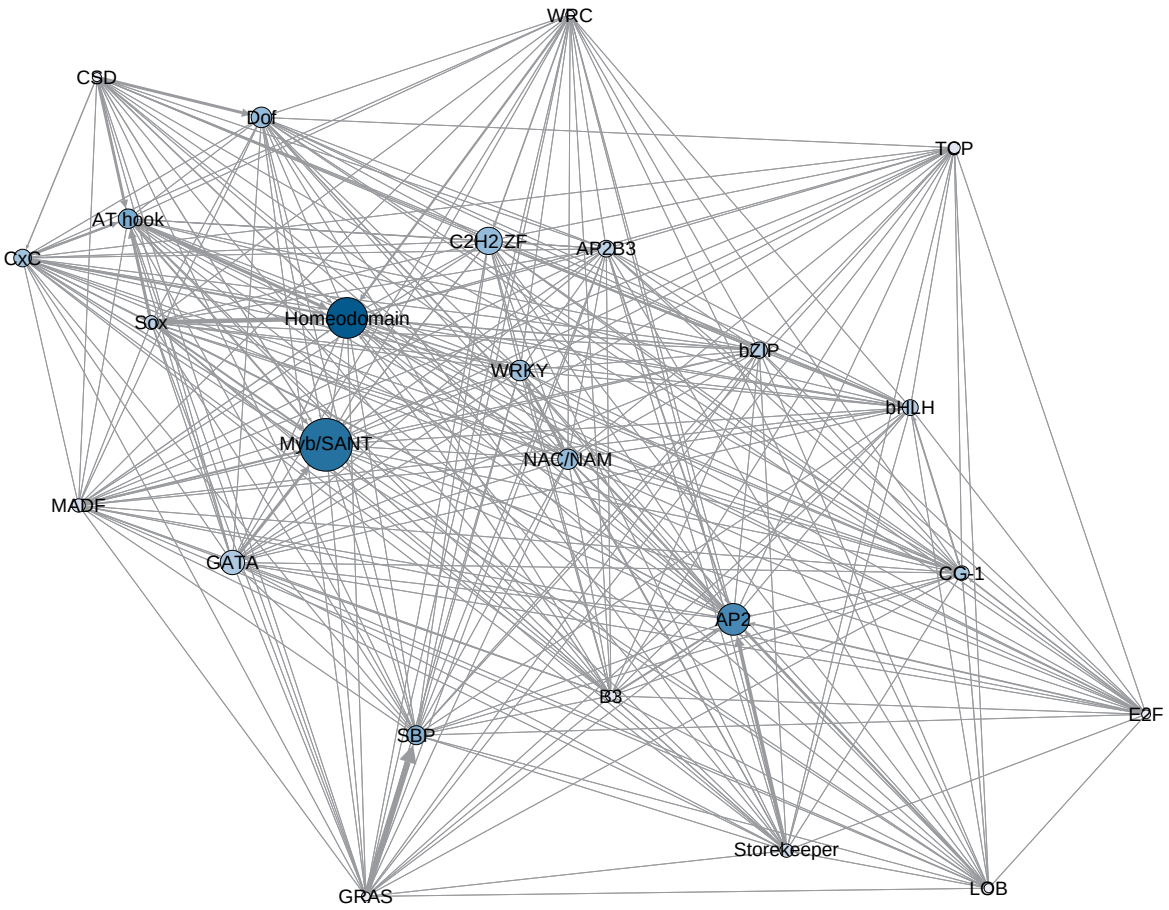
Supplementary Figure 3.13: Matrices of inter-network relationships for the genotype networks of DNA-binding domains from *A. thaliana*. Heatmaps of log10-transformed (A) overlap and (B) ϕ_{qp}, the probability of mutating from the genotype network of phenotype p to the genotype network of phenotype q. Each row and column represents a different DNA-binding domain genotype network. Domains are ordered alphabetically. Cells colored in gray indicate either N/A values (on the diagonal) or values equal to zero (off-diagonal).



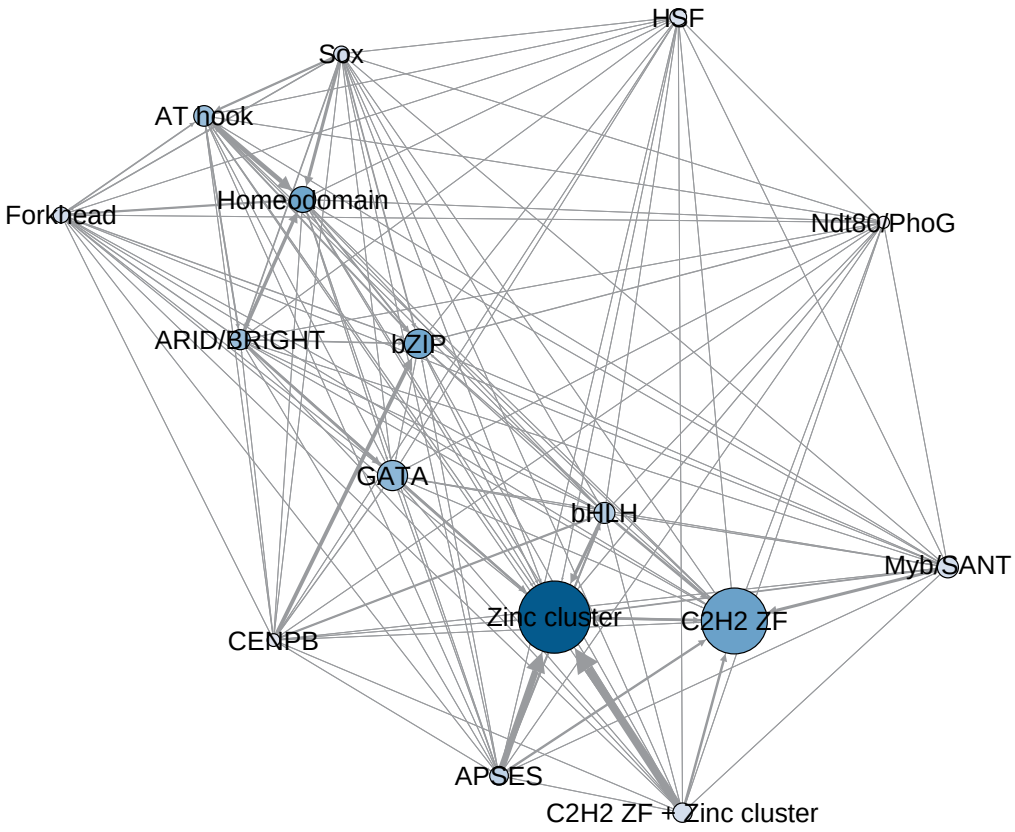
Supplementary Figure 3.14: Matrices of inter-network relationships for the genotype networks of DNA-binding domains from *N. crassa*. Heatmaps of log10-transformed (A) overlap and (B) ϕ_{qp}, the probability of mutating from the genotype network of phenotype p to the genotype network of phenotype q. Each row and column represents a different DNA-binding domain genotype network. Domains are ordered alphabetically. Cells colored in gray indicate either N/A values (on the diagonal) or values equal to zero (off-diagonal).



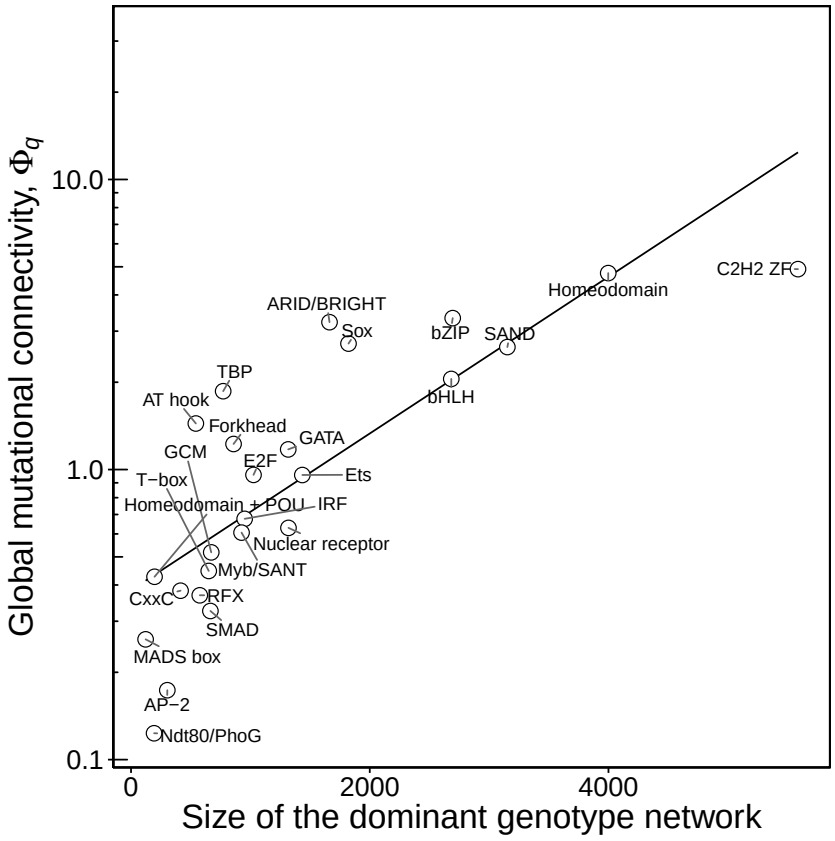
Supplementary Figure 3.15: Phenotype network for 25 DNA-binding domains from *M. musculus*. The nodes in this network represent the dominant genotype networks of DNA binding domains, and edges connect nodes if their corresponding genotype networks are connected by at least one non-neutral mutation. The size of the edges is proportional to the ϕ_{qp} among domains. Node size is proportional to the size of the associated genotype network. Node color represents the global mutational connectivity Φ_q of each domain (darker nodes have larger Φ_q).



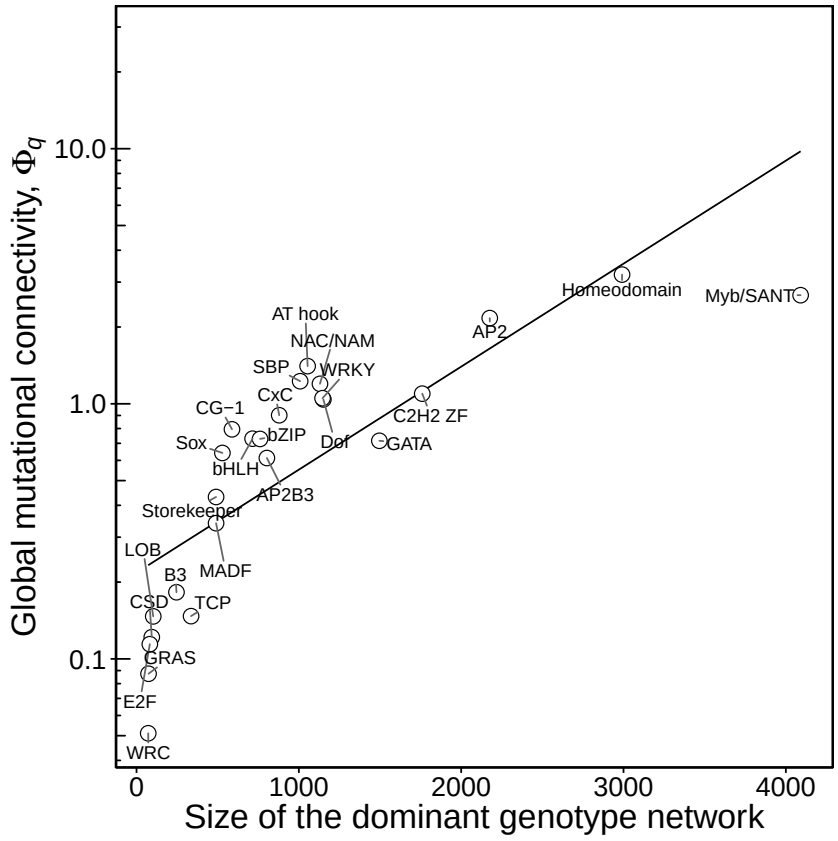
Supplementary Figure 3.16: Phenotype network for 25 DNA-binding domains from *A. thaliana*. The nodes in this network represent the dominant genotype networks of DNA binding domains, and edges connect nodes if their corresponding genotype networks are connected by at least one non-neutral mutation. The size of the edges is proportional to the ϕ_{qp} among domains. Node size is proportional to the size of the associated genotype network. Node color represents the global mutational connectivity Φ_q of each domain (darker nodes have larger Φ_q).



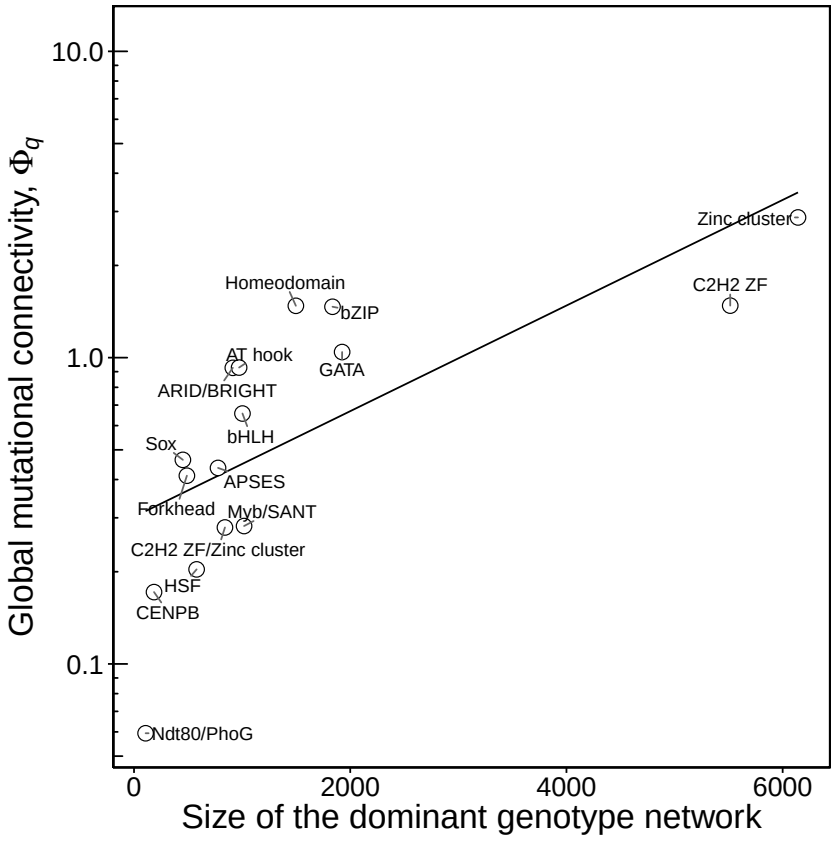
Supplementary Figure 3.17: Phenotype network for 16 DNA-binding domains from *N. crassa*. The nodes in this network represent the dominant genotype networks of DNA binding domains, and edges connect nodes if their corresponding genotype networks are connected by at least one non-neutral mutation. The size of the edges is proportional to the ϕ_{qp} among domains. Node size is proportional to the size of the associated genotype network. Node color represents the global mutational connectivity Φ_q of each domain (darker nodes have larger Φ_q).



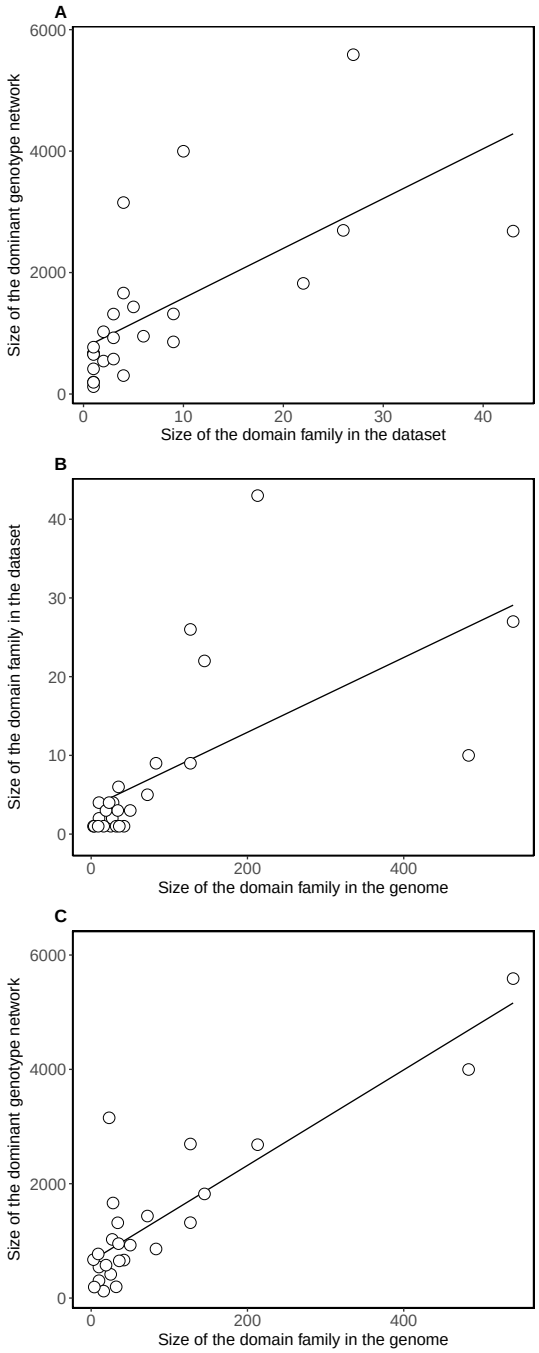
Supplementary Figure 3.18: In *M. musculus*, the global mutational connectivity of a phenotype increases with the size of its dominant genotype network. Each circle shows the global mutational connectivity Φ_q of one of the 25 *M. musculus* DNA binding domains, as a function of the number of binding sites in its dominant genotype network. The solid line is the best linear fit to the data and is provided as a visual aid.



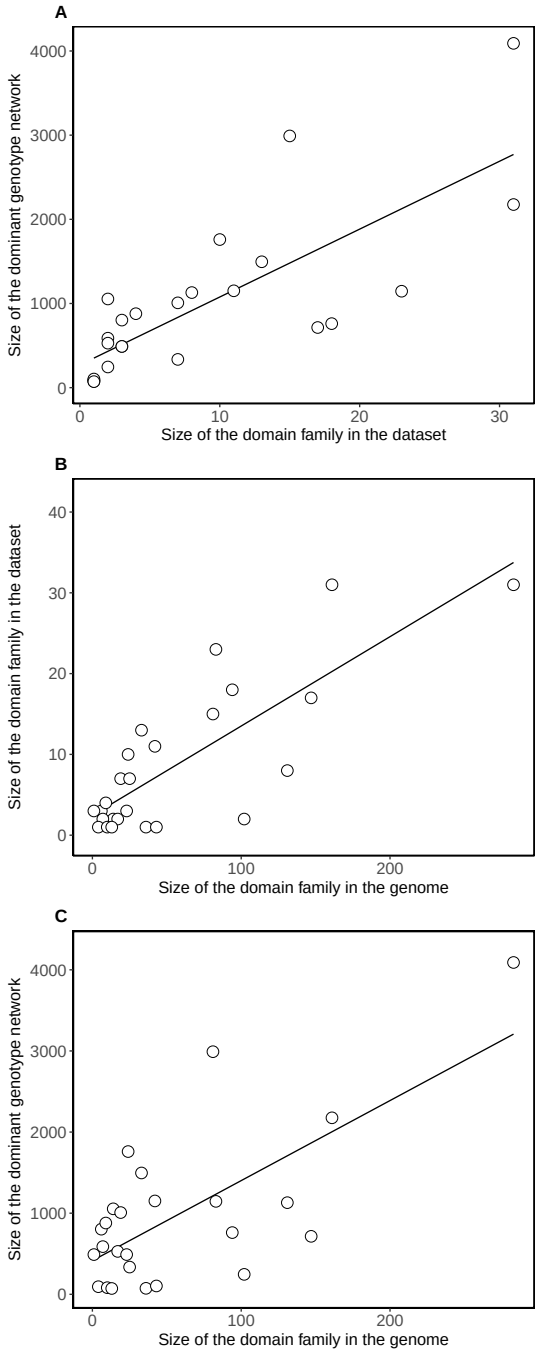
Supplementary Figure 3.19: In *A. thaliana*, **the global mutational connectivity of a phenotype increases with the size of its dominant genotype network.** Each circle shows the global mutational connectivity Φ_q of one of the 25 *A. thaliana* DNA binding domains, as a function of the number of binding sites in its dominant genotype network. The solid line is the best linear fit to the data and is provided as a visual aid.



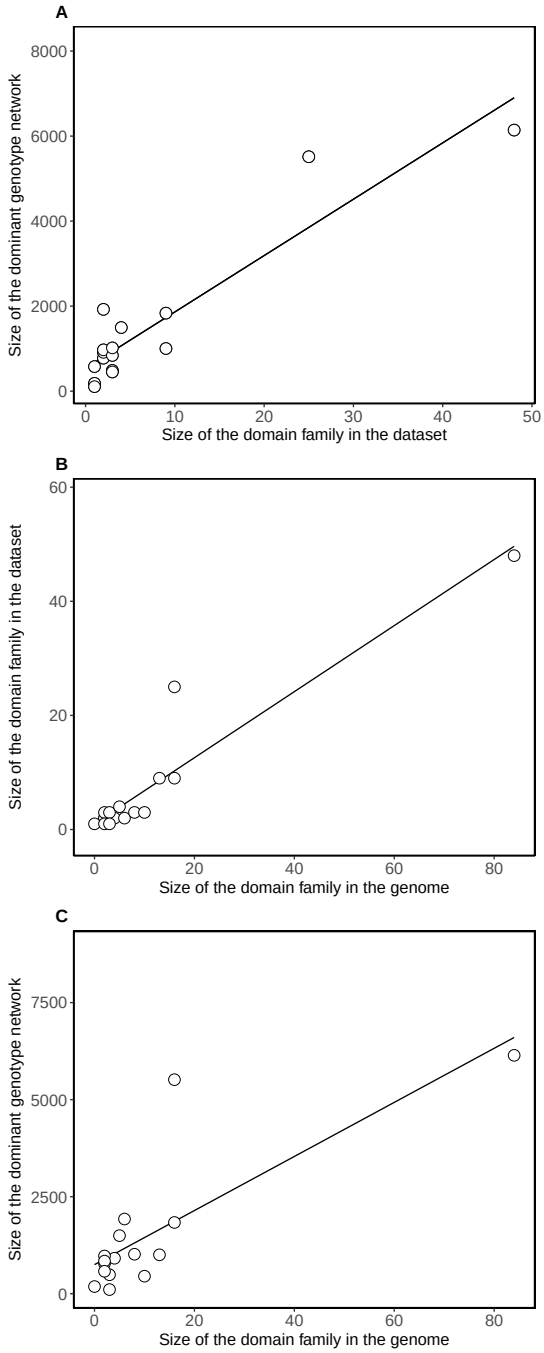
Supplementary Figure 3.20: In *N. crassa*, the global mutational connectivity of a phenotype increases with the size of its dominant genotype network. Each circle shows the global mutational connectivity Φ_q of one of the 16 *N. crassa* DNA binding domains, as a function of the number of binding sites in its dominant genotype network. The solid line is the best linear fit to the data and is provided as a visual aid.



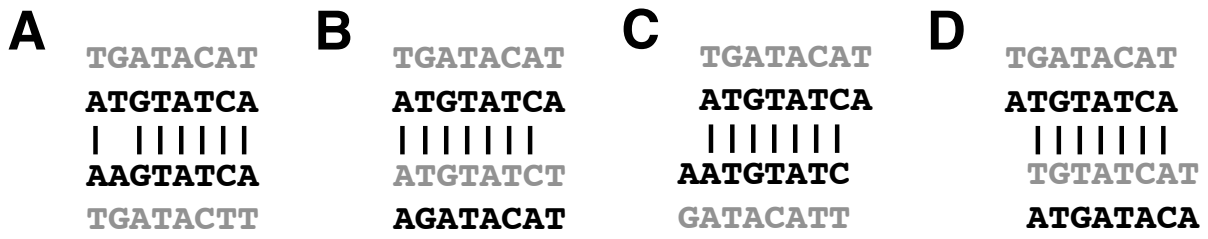
Supplementary Figure 3.21: Binding domains with more TFs have larger genotype networks in *M. musculus*. (A) The relationship between the size of a binding domain's dominant genotype network and the number of TFs per domain in our dataset (Spearman's $r = 0.8$, $p = 2 \times 10^{-6}$). (B) The relationship between the number of TFs per binding domain in our dataset and the number of TFs per binding domain in the *M. musculus* genome (Spearman's $r = 0.75$, $p = 1.4 \times 10^{-5}$). (C) The relationship between the size of a binding domain's dominant genotype network and the number of TFs per binding domain in the *M. musculus* genome (Spearman's $r = 0.7$, $p = 9.6 \times 10^{-5}$). In each panel, each circle represents one of the 25 *M. musculus* binding domains in our dataset.



Supplementary Figure 3.22: Binding domains with more TFs have larger genotype networks in *A. thaliana*. (A) The relationship between the size of a binding domain's dominant genotype network and the number of TFs per domain in our dataset (Spearman's $r = 0.83$, $p = 2.8 \times 10^{-7}$). (B) The relationship between the number of TFs per binding domain in our dataset and the number of TFs per binding domain in the *A. thaliana* genome (Spearman's $r = 0.64$, $p = 5.8 \times 10^{-4}$). (C) The relationship between the size of a binding domain's dominant genotype network and the number of TFs per binding domain in the *A. thaliana* genome (Spearman's $r = 0.44$, $p = 2.9 \times 10^{-2}$). In each panel, each circle represents one of the 25 *A. thaliana* binding domains in our dataset.



Supplementary Figure 3.23: Binding domains with more TFs have larger genotype networks in *N. crassa***.** (A) The relationship between the size of a binding domain's dominant genotype network and the number of TFs per domain in our dataset (Spearman's $r = 0.93$, $p = 2 \times 10^{-7}$). (B) The relationship between the number of TFs per binding domain in our dataset and the number of TFs per binding domain in the *N. crassa* genome (Spearman's $r = 0.94$, $p = 4.8 \times 10^{-8}$). (C) The relationship between the size of a binding domain's dominant genotype network and the number of TFs per binding domain in the *N. crassa* genome (Spearman's $r = 0.8$, $p = 2.3 \times 10^{-4}$). In each panel, each circle represents one of the 16 *N. crassa* binding domains in our dataset.



Supplementary Figure 3.24: Two forms of mutation. We consider (A,B) point mutations and (C,D) indels that shift an entire, contiguous binding site by a single base. These mutations are illustrated by aligning four different sequences with `ATGTATCA` (top bold-font sequence in each panel). Since every sequence is merged with its reverse complement (gray font), the $4^8 = 65,536$ possible sequences of length eight can be represented by a library of only 32,896 sequences. Sequences that are members of this library are represented in bold font, while their reverse complements are represented in gray font.

4 The molecular chaperone DnaK is a source of mutational robustness

Published as:

Aguilar-Rodríguez J, Sabater-Muñoz B, Montagud-Martínez R, Berlanga V, Alvarez-Ponce D, Wagner A, and Fares M A (2016) *Genome Biology & Evolution*, 8: 2979–2991.

Abstract

Molecular chaperones, also known as heat-shock proteins, refold misfolded proteins and help other proteins reach their native conformation. Thanks to these abilities, some chaperones, such as the Hsp90 protein or the chaperonin GroEL, can buffer the deleterious phenotypic effects of mutations that alter protein structure and function. Hsp70 chaperones use a chaperoning mechanism different from that of Hsp90 and GroEL, and it is not known whether they can also buffer mutations. Here, we show that they can. To this end, we performed a mutation accumulation experiment in *Escherichia coli*, followed by whole-genome resequencing. Overexpression of the Hsp70 chaperone DnaK helps cells cope with mutational load and completely avoid the extinctions we observe in lineages evolving without chaperone overproduction. Additionally, our sequence data show that DnaK overexpression increases mutational robustness, the tolerance of its clients to nonsynonymous nucleotide substitutions. We also show that this elevated mutational buffering translates into differences in evolutionary rates on intermediate and long evolutionary time scales. Specifically, we studied the evolutionary rates of DnaK clients using the genomes of *E. coli*, *Salmonella enterica*, and 83 other gamma-proteobacteria. We find that clients that interact strongly with DnaK evolve faster than weakly interacting

clients. Our results imply that all three major chaperone classes can buffer mutations and affect protein evolution. They illustrate how an individual protein like a chaperone can have a disproportionate effect on the evolution of a proteome.

4.1 Introduction

Robustness is one of the fundamental properties of living systems [51, 401–403]. This property describes the ability of a biological system to preserve its phenotype in a particular environment despite perturbations that it encounters. The robustness of a system against perturbations that are environmental (e.g., a change in temperature) is referred to as environmental robustness, whereas robustness against perturbations caused by genetic mutations receives the name of mutational or genetic robustness. Molecular chaperones [211] are one of the best-known sources of both types of robustness [403]. Chaperones, also called heat-shock proteins, assist proteins in reaching their native conformations, prevent protein aggregation, and refold misfolded proteins [213, 214, 404]. Thanks to these roles, chaperones can restore the native conformation of proteins destabilized by environmental perturbations, thus providing environmental robustness to organisms coping with stressful conditions. Because some chaperones can buffer the deleterious effects of mutations that affect protein folding, they are also a source of mutational robustness. In the context of protein evolution, chaperones are able to increase a protein's mutational robustness because they alter the mapping from protein genotypes into protein phenotypes, that is, into the structures that proteins form [212]. Specifically, they increase the number of amino acid sequences that fold into the same structure and that can perform the function associated with this structure.

There are three main chaperone systems, which are the Hsp90 system, the Hsp70 system, and the Hsp60 system (or chaperonins), of which the bacterial GroEL is a prominent member [214]. Overwhelming evidence shows that Hsp90 and GroEL can buffer mutations [216], but whether the same holds for any major chaperone from the Hsp70 system is to our knowledge unknown. A recent study has shown that RNA chaperones—they help RNA molecules to fold properly, and comprise a class of chaperones different from

these three systems—can also buffer deleterious mutations in *Escherichia coli* [405].

Pioneering work carried out by Rutherford and Lindquist [231] showed that inhibition of the chaperone Hsp90 can unveil cryptic genetic variation—genotypic variation without phenotypic variation—in the fruit fly *Drosophila melanogaster*. Subsequently, similar observations have been made in the plant *Arabidopsis thaliana* [235], the yeast *Saccharomyces cerevisae* [406] and the fish *Astyanax mexicanus* [238]. Further support was provided by Burga *et al.* [407], who found that high induction of Hsp90 during development of the nematode *Caenorhabditis elegans* reduced the penetrance of certain mutations. Additionally, Lachowiec *et al.* [408] found that paralogs of duplicated kinase-coding genes that encode a substrate of Hsp90 (i.e., a Hsp90 "client") in *S. cerevisiae* often evolve faster than paralogs encoding nonclients. In general, the rate at which nonconservative substitutions—those that alter physicochemical properties of amino acids—accumulate is especially accelerated in Hsp90 clients [409].

Multiple studies also demonstrate mutational buffering mediated by the bacterial chaperonin GroEL. For example, Fares *et al.* [258] showed that overexpressing GroEL considerably improved the fitness of *E. coli* strains with a high load of deleterious mutations, a pattern that was also observed later in *Salmonella enterica* [267]. Moreover, GroEL overexpression in *E. coli* increases the ability of GroEL client proteins to tolerate mutations [266, 278, 410, 411], as well as their ability to undergo adaptive evolution [278, 411]. Buffering of destabilizing mutations accelerates the evolutionary rates of GroEL clients [279, 280, 409, 412].

While no Hsp70 chaperone has been directly implicated in mutational buffering, pertinent circumstantial evidence exists. For example, DnaK—the major bacterial Hsp70 chaperone—is overexpressed together with GroEL in *S. enterica* lineages with reduced fitness caused by the accumulation of deleterious mutations [267]. In addition, *D. melanogaster* populations showing inbreeding depression, where increased homozygosity exposes recessive deleterious mutations, significantly up-regulate the expression of Hsp70 compared with outbred populations [413].

The chaperones from the Hsp70 system are very conserved from bacteria to humans [414]. They play a central role in proteome integrity, and are involved both in co- and post-translational folding [214]. In bacteria, the Hsp70 chaperone DnaK (together with GroEL and the Trigger Factor) is one of the main molecular chaperones, where it is the central hub in the chaperone network of the cytosol [240, 241]. It interacts with at least 700 mostly cytosolic proteins [240]. The DnaK interactome was characterized by the isolation of DnaK interactors using immobilized metal affinity chromatography, followed by liquid chromatography mass spectrometry. These regular clients of DnaK are enriched for proteins with low intrinsic solubility, proteins that tend to be members of hetero-oligomeric complexes and/or proteins that show a high density of hydrophobic patches flanked by positive residues [240]. DnaK is highly expressed constitutively and essential at 42 °C [240, 241]. During its ATP-dependent reaction cycle, DnaK interacts with the Hsp40 co-chaperone DnaJ, which determines the client binding specificity of DnaK [244, 245], and the nucleotide exchange factor GrpE [214]. The chaperone system formed by these three proteins can both fold nascent proteins and refold denatured proteins. It does so by binding to exposed hydrophobic patches in unfolded or partially folded protein substrates, thus preventing detrimental interactions with other polypeptides in the crowded cellular milieu. By successively binding and releasing a protein substrate in a cyclic process that consumes ATP, the chaperone system DnaK–DnaJ–GrpE allows the substrate to gradually explore its complex folding energy landscape [213, 214]. For some proteins (~20% of the total proteome), several of these bind-release cycles are enough to achieve the native conformation. However, other proteins (~10% of the total proteome) still require the downstream chaperone system GroEL/ES [214]. The importance of the DnaK–DnaJ–GrpE system in the bacterial chaperone network is obvious from its strong conservation across bacteria, except for two species from the order Aquificales that have lost the entire system, and individual losses of *dnaJ* and *grpE* in obligate endosymbionts that have experienced considerable genome reductions [415].

Most mutations affecting proteins are neutral or deleterious [416], and functionally

important mutations often destabilize proteins [276, 411]. If DnaK buffers destabilizing mutations, then the deleterious effects of mutations in highly interacting (strong) clients should be lower than in sporadic (weak) clients, where they should be lower than in nonclients. In other words, the more strongly a protein's integrity depends on DnaK, the higher should be its tolerance to mutations, and the lower the signature of purifying selection that purges those mutations. With this reasoning in mind, we here use laboratory experiments to evaluate the effect of DnaK buffering on the evolution of its client proteome on short evolutionary time scales. We complement our experimental observations with sequence analyses to study the effect of DnaK on intermediate and long evolutionary time scales.

4.2 Results

4.2.1 Experimental evolution of *E. coli* under DnaK overexpression

To study the effect of DnaK overexpression on protein evolution experimentally, we performed mutation accumulation experiments similar to those we reported recently for the chaperonin GroEL, but for DnaK overexpression [266]. Briefly, we initiated 68 parallel and independent clonal lines of evolution, all of which derived from the same hypermutable clone (*E. coli* K12 MG1655 Δ*mutS*) [266] (fig. 4.1A). Cells of the 68 lines all harbored the plasmid pKJE7, which contains the operon *dnaK–dnaJ–grpE* under the control of the L-arabinose-inducible *araB* promoter P_{BAD} [417]. We refer to this strain as $DnaK^+$. We evolved 60 of the 68 $DnaK^+$ lines through repeated single-cell bottlenecks in the presence of the inducer, to ensure overexpression of DnaK, as well as of the cochaperone DnaJ and the nucleotide exchange factor GrpE. All evolving lineages were passaged after 24 hours of incubation. Because of the bottlenecks to which we exposed the populations, genetic drift was strong and the efficiency of selection was weak during the experiment, such that nonlethal mutations are free to accumulate [44]. We evolved 30 of the 60 clonal lines at 37 °C, and the other 30 at 42 °C. The higher temperature serves to increase the

deleterious effect of destabilizing mutations in the bacterial proteome [241]. Finally, the remaining 8 DnaK$^+$ lines were evolved in the absence of inducer, and therefore without DnaK overexpression (4 lines at 37 °C and the other 4 at 42 °C).

At each of the two temperatures, we additionally evolved 8 control clonal lines founded from the same parental strain, but carrying a pKJE7-derived plasmid where the operon *dnaK–dnaJ–grpE* is deleted (fig. 4.1B). Cells of all 16 control lines therefore cannot overexpress DnaK, even though their growth medium contains L-arabinose (DnaK$^-$ lines). At each temperature, half of the lines evolved in the presence of L-arabinose, whereas the other half evolved in medium devoid of this expression inducer. In total, we evolved 86 bacterial populations: 68 DnaK$^+$ lines and 16 DnaK$^-$. We stopped the evolution experiment after 85 single-cell bottlenecks, or 1,870 generations (assuming conservatively ~22 generations per daily growth cycle).

4.2.2 Evolving lineages tend to go extinct in the absence of DnaK overexpression

One of the first indications that DnaK overexpression could be buffering deleterious mutations accumulated during the evolution experiment is the observed pattern of extinctions (fig. 4.1). Some evolving lines went extinct, presumably due to high levels of mutational load, and remarkably, all extinctions occurred in lines that were not overexpressing DnaK. They were either DnaK$^-$ lines or DnaK$^+$ evolved in the absence of the inducer. More specifically, 75% of the DnaK$^-$ lines (12 of 16) went extinct before the end of the evolution experiment (fig. 4.1B). Among the DnaK$^+$ lines, 62.5% of the lines (5 of 8) evolving in the absence of the inducer went extinct, whereas none of the 60 lines evolving in the presence of the inducer experienced any extinction (fig. 4.1A). This observation strongly suggests that overexpressing the chaperone DnaK has increased the robustness of the cells to the accumulation of deleterious mutations, helping them cope with mutational load.

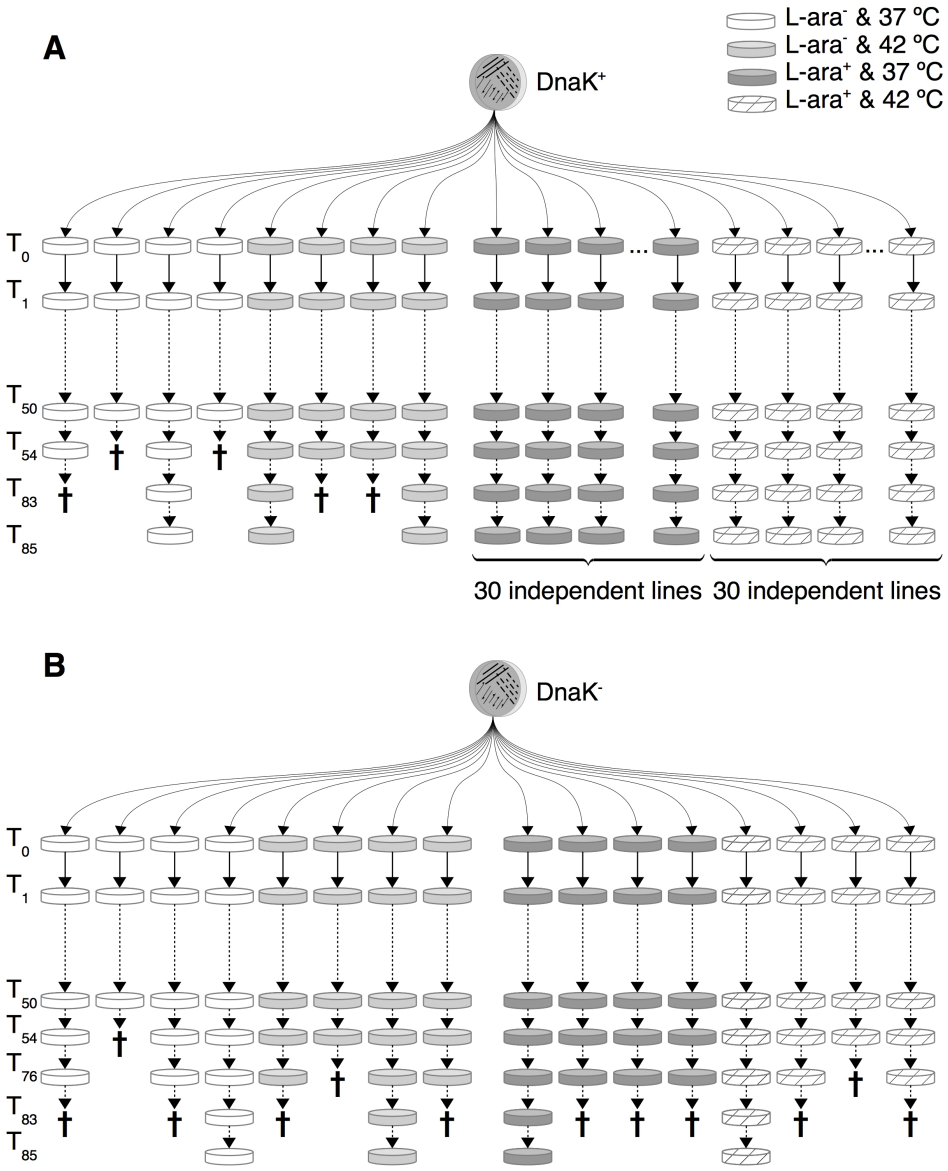


Figure 4.1: Mutation accumulation experiment. Evolutionary history of the populations evolved in this study from the first daily transfer or single-cell bottleneck (T_0) until the end of the evolution experiment (T_{85}). We constructed two strains derived from an ancestral *E. coli* K-12 MG1655 strain lacking the mismatch repair gene *mutS*. The DnaK$^+$ strain harbours the 15-copy plasmid pKJE7 that contains the DnaK/DnaJ/GrpE chaperone system under the control of the promoter P_{BAD} inducible by L-arabinose [417]. The DnaK$^-$ strain contains a control pKJE7-derived plasmid where the operon *dnaK–dnaJ–grpE* has been deleted. We evolved in parallel multiple independent populations of both strains through single-cell bottlenecks under the effect of strong genetic drift at two different temperatures (37 °C and 42 °C). At each temperature we evolved some populations in the presence of L-arabinose (L-ara$^+$), and some in the absence of this expression inducer (L-ara$^-$). (**A**) During the evolution of 68 DnaK$^+$ populations, five out of eight lines evolving in the absence of inducer went extinct (indicated by a cross). None of the 60 lines evolving under DnaK overexpression experienced any extinction. (**B**) Of the 16 independent DnaK populations, 12 populations went extinct. We finished the evolution experiment after 85 single-cell bottlenecks (T_{85}), or ~1,870 generations.

4.2.3 Overexpressing DnaK increases the robustness to nonsynonymous mutations of DnaK clients

In order to study the effect of DnaK buffering on genome evolution, we sequenced the genomes of some lines at the end of the evolution experiment, after 85 passages, and compared them to the ancestral genome, which we had sequenced in a previous study [266]. Among the clonal lines evolved in the presence of the inducer L-arabinose, we randomly selected for sequencing 3 DnaK$^+$ lines evolved at 37 °C, and another 3 at 42 °C. We also sequenced the only two surviving control DnaK$^-$ lines evolved with L-arabinose in the medium, each at a different temperature. Although all sequenced lines evolved in the presence of the inducer, only DnaK$^+$ lines are able to overexpress the chaperone.

In order to evaluate if a significant difference existed in the mutation rate (or generation time) between the sequenced DnaK$^+$ and DnaK$^-$ lines, we compared the number of accumulated synonymous mutations between them (supplementary table 4.1). We observed an average number of 78 synonymous mutations per DnaK$^+$ line and 66 synonymous substitutions per DnaK$^-$ line, which is not significantly different (binomial test, $P = 0.359$; supplementary table 4.1). We also did not observe any significant difference in the number of accumulated nonsynonymous mutations (binomial test, $P = 0.646$; table 4.1), the number of indels (binomial test, $P = 0.332$; supplementary table 4.1), or the ratio of transitions to transversions (χ^2 test, $P = 0.273$; supplementary table 4.1).

We also verified that DnaK was still overexpressed at the end of the experiment in the 8 sequenced lines. The overexpression of DnaK may be energetically costly, just as is the case for the chaperonin GroEL [258, 266]. In principle, this cost could favor the accumulation of mutations that lead to a decrease in the expression of DnaK during the evolution experiment, especially if the energetic cost of overproducing the chaperone is greater than the benefits derived from mutational buffering [266]. However, we observed that for the sequenced lines the overexpression of DnaK was maintained through the mutation accumulation experiment at both 37 °C and 42 °C (fig. 4.2; supplementary fig. 4.1). In the presence of the inducer L-arabinose, all DnaK$^+$ lines overexpressed DnaK not only

at the start of the evolution experiment, but also at the end, except for one of the DnaK$^+$ lines evolved at 42 °C. However, this loss of overexpression occurred towards the end of the experiment and even then DnaK was still overexpressed for most of the daily growth cycle of this line (supplementary fig. 4.2). In no line did we observe overexpression in the absence of the inducer. The control DnaK$^-$ lines always exhibited wild-type expression levels of DnaK.

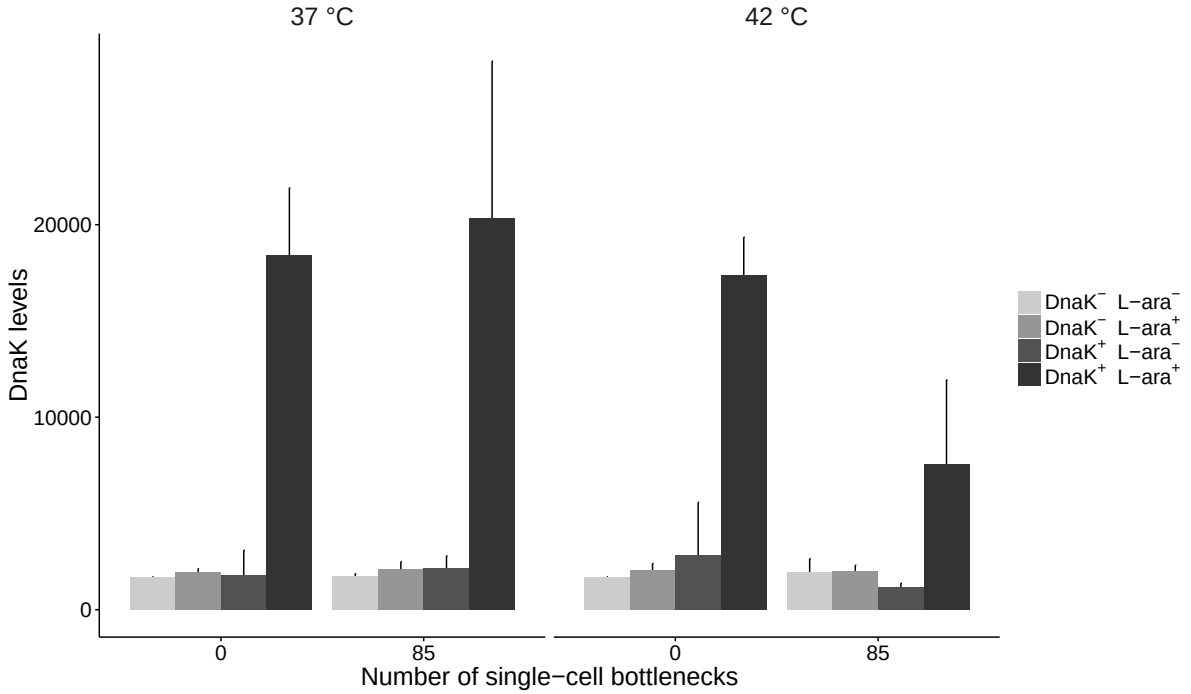


Figure 4.2: DnaK abundance at the beginning and the end of the mutation accumulation experiment. We measured the abundance of the chaperone DnaK for the 8 sequenced lines evolved trough 85 single-cell bottlenecks (~1,870 generations) at 37 °C or 42 °C. For comparison, we also measured the abundance of the chaperone in the ancestral DnaK$^+$ and DnaK$^-$ strains at both temperatures. We determined DnaK levels in the presence and absence of the inducer L-arabinose (L-ara$^+$ and L-ara$^-$, respectively), as described in Materials and Methods ("Verification of DnaK overexpression"), via the intensity of the DnaK band in a Western blot. The evolved lines did not lose the ability to overexpress DnaK in the presence of the inducer L-arabinose except for a DnaK$^+$ line evolved at 42 °C (line #2), which explains the decrease in the average DnaK abundance at the end of the evolution experiment. However, this loss of overexpression occurred late in the evolution experiment, and it is not even complete for most of the daily growth cycle of this line (supplementary fig. 4.2). The height of the bars indicates mean DnaK abundance across two experimental replicates per strain and condition. Error bars represent 1 SD of the mean.

In the genomes of the evolved DnaK$^+$ lines, we first studied the incidence of nonsyn-

Table 4.1: Distribution of nonsynonymous nucleotide substitutions among DnaK client and nonclient proteins after ~1,870 generations of evolution in mutation accumulation experiments conducted at 37 °C and 42 °C

Temperature	Line[a]	Number of mutations			
		Clients			Nonclients
		Strong[b]	Weak[c]	Total	
37 °C	$DnaK^+$ #1	3	0	12	50
	$DnaK^+$ #2	7	2	19	91
	$DnaK^+$ #3	11	2	25	91
	$DnaK^-$	12	1	17	108
42 °C	$DnaK^+$ #1	11	1	22	95
	$DnaK^+$ #2	13	3	27	103
	$DnaK^+$ #3	9	2	22	115
	$DnaK^-$	11	1	15	99

[a] Experimental evolution lines sequenced in this study. For each temperature, we sequenced three lines overexpressing the DnaK–DnaJ–GrpE chaperone system ($DnaK^+$ lines) and a control line where this system is expressed at wild type levels ($DnaK^-$ line).

[b] Strong clients are those with a high relative enrichment factor on DnaK within the third quartile of the distribution.

[c] Weak clients are those with a low relative enrichment factor on DnaK within the first quartile of the distribution.

onymous nucleotide substitutions among DnaK clients and nonclients (table 4.1). In this analysis, we considered as nonclients all proteins from the *E. coli* proteome that are not part of a set of 674 DnaK clients determined by Calloni *et al.* [240], and analyzed the lines evolved at 37 °C and 42 °C independently. To improve statistical power, we combined mutations across $DnaK^+$ lines evolved at the same temperature after classifying them according to whether they affect DnaK clients or nonclients. If DnaK is buffering deleterious mutations, we would expect a higher proportion of mutations affecting clients in the lines evolved under DnaK overexpression.

In the $DnaK^-$ line evolved at 37 °C, ~14% of nonsynonymous mutations (17 out of 125) affected DnaK clients. Compared with this proportion when DnaK is not overexpressed, the proportion of mutations in clients in the $DnaK^+$ lines was significantly higher (56 out of the total 288 mutations, ~19%; binomial test: $P = 0.006$; fig. 4.3A). Similarly, compared with the DnaK line evolved at 42 °C, where ~13% of all mutations affected DnaK clients (15 out of 114 mutations), the $DnaK^+$ lines showed significantly more mutations in clients

(71 out of 384 total mutations, ~18%; binomial test: $P = 0.003$; fig. 4.3A). These results suggest that overexpressing DnaK does indeed increase the robustness of its clients to amino acid replacements. Temperature itself had no significant effect on the fraction of all mutations affecting DnaK clients in DnaK$^+$ lines (Fisher's exact test: odds ratio $F = 1.06$, $P = 0.77$) and DnaK$^-$ lines (Fisher's exact test: odds ratio $F = 1.04$, $P = 0.999$).

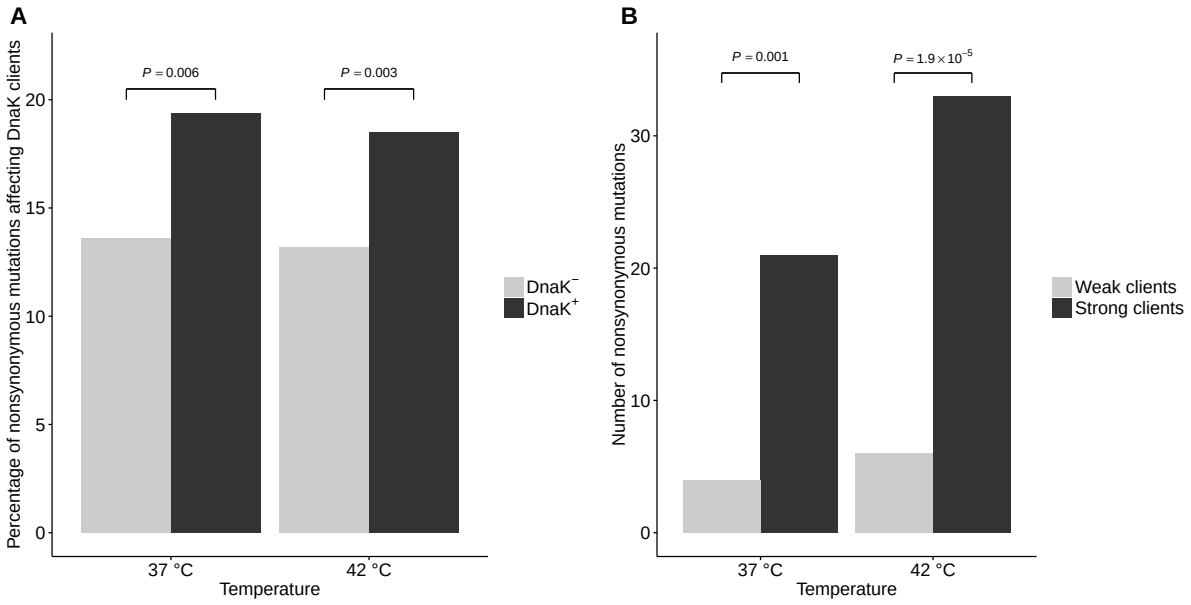


Figure 4.3: Nonsynonymous mutations accumulated in DnaK clients. (**A**) The proportion of nonsynonymous mutations that affect DnaK clients is significantly higher in DnaK$^+$ lines that overexpress DnaK than in the control DnaK$^-$ lines that do not express the chaperone at such high levels. This is observed both for lines evolved at 37 °C and 42 °C. We combined mutations across DnaK$^+$ lines evolved at the same temperature. The significance of the difference in the proportions was evaluated using a binomial test. (**B**) At both temperatures, strong clients have accumulated significantly more nonsynonymous substitutions than weak clients in DnaK$^+$ lines. Strong clients include those clients with the highest DnaK dependency, whereas weak clients include clients with the lowest chaperone dependency. Statistical significance was evaluated using Fisher's test.

4.2.4 Strong DnaK clients accumulate more nonsynonymous mutations than weak clients

Next, we studied if strongly interacting DnaK clients are more robust to mutations than weakly interacting clients, as evidenced by the pattern of mutations fixed in the mutation accumulation experiment under DnaK overexpression. To assess how strongly a protein

depends on DnaK for folding, we used recent experimental proteomic data which determined how strongly 668 DnaK-interacting proteins interact with DnaK by measuring the fraction of cellular protein bound to DnaK at 37 °C, a property that correlates with chaperone dependency for folding and maintenance and residence time of the protein on DnaK [240]. In a Δ*dnaK E. coli* strain, strong clients are more prone to form aggregates than weak clients, indicating that the relative enrichment of a protein on DnaK is a good proxy for the dependence upon DnaK for folding [240]. We consider as strong clients those with a relative enrichment factor on DnaK within the third quartile of the distribution of DnaK dependency (N=167), and weak clients those within the first quartile (N=167). Therefore, strong clients include those clients with the highest DnaK dependency, whereas weak clients include clients with the lowest chaperone dependency.

In the DnaK$^+$ lines evolved at 37 °C, we found more nonsynonymous substitutions in strong clients (21 mutations) than in weak clients (4 mutations) (table 4.1). Considering the number of nonsynonymous sites in strong clients (43,731 sites) and weak clients (41,238 sites), this difference was significant (Fisher's exact test: F=4.95, P=0.001; fig. 4.3B). At 42 °C, the results were similar, with 33 mutations in strong clients and 6 in weak clients (Fisher's exact test: F=5.12, $P = 1.9 \times 10^{-5}$; fig. 4.3B; table 4.1). In conclusion, at both temperatures, client proteins that are more dependent upon DnaK for folding accumulate significantly more mutations than less dependent clients.

4.2.5 DnaK accelerates protein evolution on intermediate and long evolutionary time scales

We wanted to find out if the DnaK-mediated mutational buffering we observed on the short time scales of laboratory evolution has also left signatures on longer evolutionary time scales. To this end, we determined two measures of evolutionary rates for protein-coding genes from gamma-proteobacteria. The first, nonsynonymous divergence among one-to-one orthologs of *E. coli* and *S. enterica*, is relevant for intermediate evolutionary time scales. The second, protein (amino acid) distance among orthologous proteins found

in 85 gamma-proteobacterial genomes (including *E. coli* and *S. enterica*), is relevant for long time scales. We employ protein distance instead of nonsynonymous distance because amino acid replacements are less sensitive than nucleotide substitutions to the expected loss of phylogenetic signal between sequences of distantly related taxa. To assess how strongly a protein depends on DnaK for folding, we use the relative enrichment of the protein on DnaK as a proxy for the dependence of the protein upon DnaK for folding [240]. We note that this interaction strength is more likely to have remained unchanged during the divergence of *E. coli* and *S. enterica*, than during the divergence of all the other 83 gamma-proteobacterial species we analyzed.

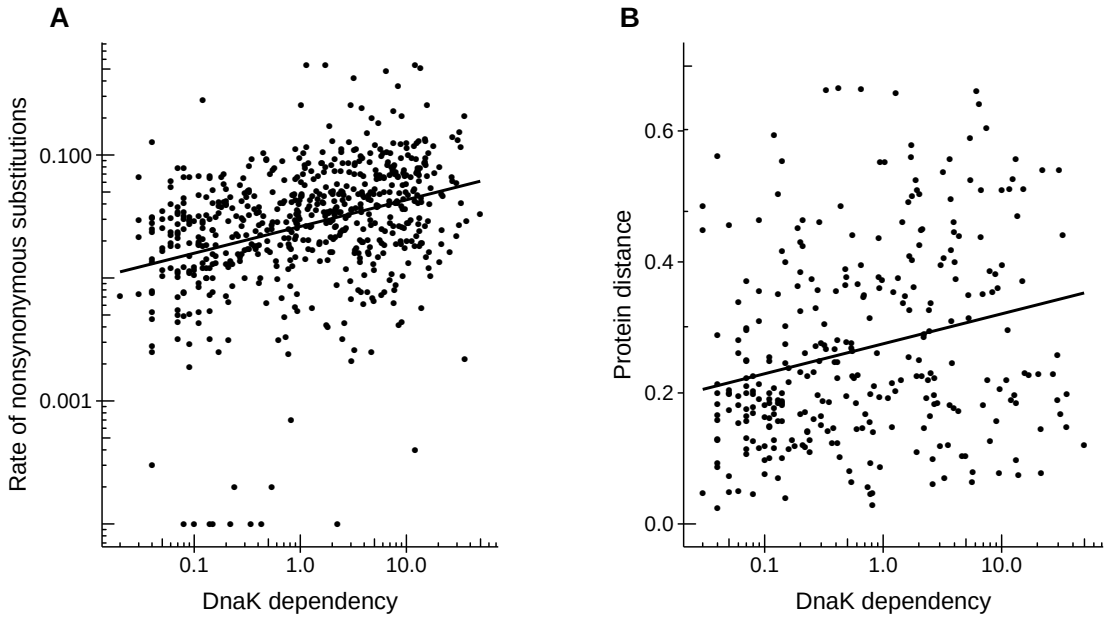


Figure 4.4: DnaK accelerates protein evolution on intermediate and long evolutionary time scales. Scatter-plots showing the relationship between DnaK dependency (calculated as a relative enrichment factor that indicates the fraction of cellular protein bound to DnaK at 37 °C, horizontal axis) and the degree of divergence over (**A**) intermediate time scales, measured as nonsynonymous divergence (Spearman rank correlation coefficient, $\rho = 0.367$, $N = 627$, $P < 2.2 \times 10^{-16}$), and (**B**) long time scales, measured as protein (amino acid) distance ($\rho = 0.257$, $N = 311$, $P = 4.4 \times 10^{-6}$) (vertical axes). Solid lines represent the best fit to the points. Note the logarithmic scale on both axes.

We find a strong and highly significant positive association between DnaK dependency and the rate of nonsynonymous substitutions for *S. enterica* and *E. coli* (Spearman's rank

correlation coefficient, $\rho = 0.367$, $N = 627$, $P < 2.2 \times 10^{-16}$; fig. 4.4A). This indicates that the stronger the interaction of a protein with DnaK, the faster the protein evolves. The same pattern is obtained at the larger time scales of protein distances for 85 gamma-proteobacterial genomes ($\rho = 0.257$, $N = 311$, $P = 4.4 \times 10^{-6}$; fig.4.4B). Gene expression level, which is the most important determinant of protein evolutionary rates, at least in unicellular organisms [222, 224, 385], is a possible confounding factor in this analysis. For example, using codon usage bias (CUB) as a proxy for gene expression, we observe that genes with higher CUB show lower nonsynonymous divergence ($\rho = 0.558$, $N = 1014$, $P < 2.2 \times 10^{-16}$), protein distance ($\rho = 0.255$, $N = 3159$, $P < 2.2 \times 10^{-16}$) and DnaK dependency ($\rho = 0.262$, $N = 627$, $P = 2.5 \times 10^{-11}$). However, the association between DnaK dependency and evolutionary rate cannot be solely explained by this confounding factor: A partial correlation analysis shows that the association still holds after controlling for CUB, both on intermediate time scales ($\rho = 0.295$, $N = 627$, $P = 1.2 \times 10^{-14}$) and long time scales ($\rho = 0.229$, $N = 311$, $P = 3.8 \times 10^{-5}$). We use CUB instead of gene expression data here for two main reasons. First, we can compute CUB for all 631 DnaK clients in our data set, whereas expression data is only available for 457 clients. Second, gene expression data has been measured in just one environment and one strain of *E. coli*, whereas CUB is the result of selective pressures imposed by many different environments over long periods of time. Nonetheless, the association between evolutionary rate and DnaK dependency still holds after correcting for gene expression directly (supplementary information section 4.5.1). Together, these results indicate that the chaperone DnaK affects protein evolution in accordance with the mutational buffering hypothesis. Importantly, this effect is not only independent of CUB and gene expression, but also of other biological factors, such as essentiality and number of protein-protein interactions (supplementary information section 4.5.2 and table 4.2).

In a subsequent analysis, we find that clients evolve more slowly than nonclients (supplementary fig. 4.3 and supplementary information section 4.5.3). This last difference cannot be explained by the number of protein-protein interactions, by essentiality, or by

CUB as confounding factors (supplementary information section 4.5.3 and supplementary tables 4.3 and 4.3). The reason for this observation could be that clients are intrinsically less robust to mutations than nonclients due to some general physicochemical difference. For example, Calloni *et al.* [240] found that DnaK clients have generally low solubility, often belong to heterooligomeric complexes, and are prone to misfolding. However, in accordance with the mutational buffering hypothesis we observe that strong clients evolve faster than weak clients (fig. 4.5; supplementary information section 4.5.3). The accelerated evolution of strong clients compared with weak clients exactly mirrors the greater accumulation of nonsynonymous mutations in strong clients during the evolution experiment (fig. 4.4B).

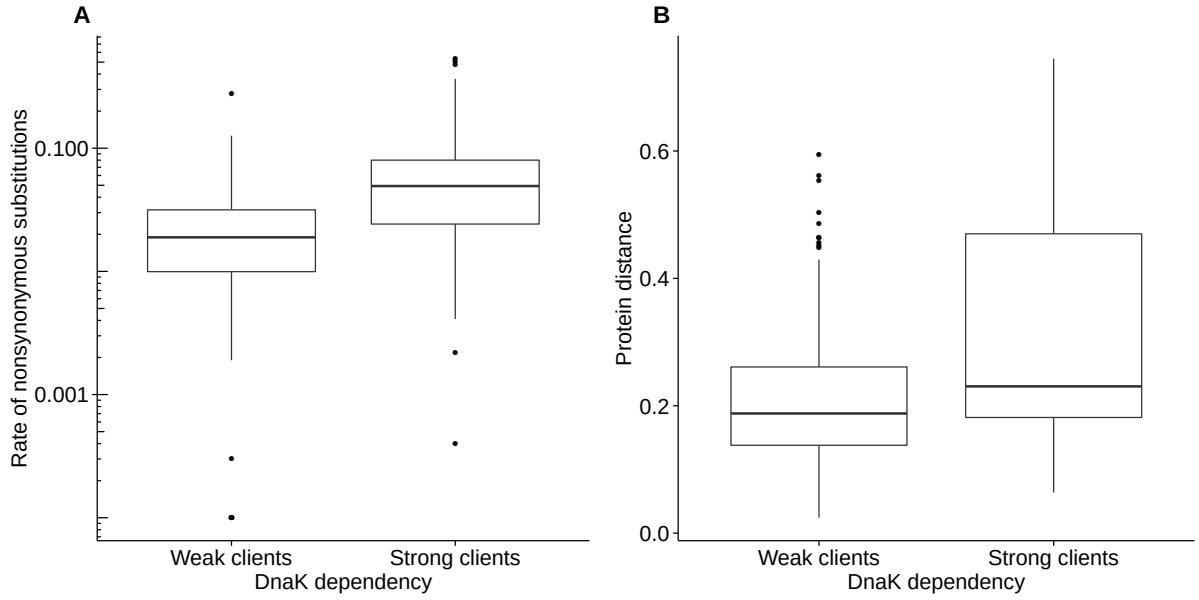


Figure 4.5: Strong clients evolve faster than weak clients. (**A**) We find that strong clients evolve faster than weak clients on intermediate evolutionary time scales, measured as the rate of nonsynonymous substitutions (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$). (**B**) On long evolutionary time scales, we also find that strong clients evolve faster than weak clients (Wilcoxon rank-sum test, $P = 2.3 \times 10^{-3}$). The thick horizontal line in the middle of each box represents the median of the data, whereas the bottom and top of each box represent the 25^{th} and 75^{th} percentiles, respectively. Note the logarithmic scale on the y axis in (**A**).

4.2.6 DnaK-mediated acceleration of protein evolution is independent of GroEL buffering

The ability of DnaK to facilitate the accumulation of nonsynonymous mutations in DnaK clients resembles the well-studied mutational buffering by the chaperonin GroEL [258, 266, 278, 410, 411]. Additionally, the observed correlation between DnaK dependency and protein evolutionary rates is similar to the previously reported acceleration of protein evolution by GroEL [279, 280]. We therefore removed known GroEL clients from our data set to investigate if our observations are independent of the effect of GroEL on mutation accumulation and evolutionary rates. We defined the GroEL interactome in *E. coli* as the union of two previously reported sets of GroEL interactors [249, 250]. Of the 253 GroEL clients that comprise the GroEL interactome, there are 122 proteins that are also clients of DnaK.

The observation that DnaK overexpression increases the proportion of nonsynonymous substitutions affecting DnaK clients is still significant after removing GroEL clients. Combining mutations from the $DnaK^+$ lines evolved at 37 °C and 42 °C we find that ~16% of mutations (97 out of 621) affected DnaK clients, which is significantly higher than what we find in the $DnaK^-$ lines (28 out of 230 mutations, ~12%; binomial test: $P = 0.01$). Similarly, considering the number of nonsynonymous sites in strong clients (36,026 sites) and weak clients (33,047 sites) after removing GroEL clients, we still find that strong DnaK clients accumulate more nonsynonymous substitutions than weak clients (Fisher's exact test: odds ratio $F = 4.7$, $P = 2 \times 10^{-5}$). Finally, the positive association between DnaK dependency and evolutionary rates still holds after removing GroEL clients and controlling for CUB in a partial correlation analysis, both on intermediate time scales ($\rho = 0.318$, $N = 511$, $P = 3.9 \times 10^{-14}$) and long time scales ($\rho = 0.226$, $N = 240$, $P = 3.6 \times 10^{-4}$).

4.3 Discussion

We show how the overexpression of the DnaK–DnaJ–GrpE chaperone system over the course of a mutation accumulation experiment increases the proportion of nonsynonymous substitutions affecting DnaK clients. In addition, strong clients accumulate more nonsynonymous mutations than weak clients. Additional evidence of mutational buffering by DnaK is provided by the observation that evolving lines overproducing this chaperone avoid extinction after experiencing 85 single-cell bottlenecks. Recently, we obtained similar results in hypermutable *E. coli* cells evolving in identical conditions but overproducing the GroEL-GroES chaperonin system [266]. There, we observed that lines evolving with high levels of GroEL were not only less prone to extinction under strong genetic drift than control lines, but also that they were accumulating significantly more indels and replacements between amino acids belonging to different physicochemical categories.

We also find that DnaK-mediated mutational buffering has left a trace in DnaK clients during the divergence of 85 different gamma-proteobacterial species over much longer evolutionary time scales than those explored in our laboratory evolution experiment. We find that clients that depend more on DnaK for folding tend to evolve faster than less interacting clients. Similar chaperone-mediated accelerations of protein evolution have been observed in GroEL clients [279, 280] and Hsp90 clients [408, 409]. However, we notice that DnaK clients evolve slower than proteins not known to be DnaK interactors [240]. This is likely the result of important physicochemical differences between clients and nonclients. For example, clients are prone to aggregation and misfolding [240], which may make them intrinsically less robust to destabilizing mutations.

Despite the great differences in the mechanism of chaperone action between the three major chaperone families—chaperonins, Hsp90 chaperones and Hsp70 chaperones—[214, 216, 404, 418], at least some of their members seem to have qualitatively comparable effects on protein evolution. But protein chaperones are not the only chaperones that can increase the mutational robustness of their substrates: A recent study has found that some RNA chaperones can buffer deleterious mutations in *E. coli* and therefore

affect RNA evolution [405]. These chaperones, which are completely unrelated to protein chaperones, are RNA-binding proteins that facilitate the proper folding of RNA molecules. Elucidating to what extent the buffering mechanisms of all these chaperones differ is an important future direction of enquiry.

Thanks to their fostering of mutational robustness, chaperones can facilitate evolutionary innovations [212], even though we do not study such innovations here. The increase in the mutational robustness of a protein caused by chaperone interactions reduces the efficiency of purifying selection in purging mutations in the protein. Thanks to chaperone-mediated buffering, many such mutations are neutral and can persist in a population. Importantly, these cryptic genetic variants may include preadaptive mutations that can generate evolutionary innovations in new environments [278, 411]. To illuminate if and how DnaK can increase the ability to evolve functional innovations of its client proteome will also be an interesting subject for future work.

In summary, we analyzed the evolution of proteins that are subject to DnaK-assisted folding on short, intermediate, and long evolutionary time scales through a combination of experimental and comparative approaches. Most of our evidence indicates that the bacterial chaperone DnaK can buffer mutations in its client proteins, and that these proteins therefore evolve faster than in the absence of DnaK-mediated folding. This is, to our knowledge, the first demonstration that a member of the Hsp70 family can buffer the effect of mutations, with long-term consequences on protein evolution [216]. Through its role in protein folding, an individual chaperone such as DnaK can have a disproportionate effect on proteome evolution, and thus on genome evolution.

4.4 Material and methods

4.4.1 Bacterial strains and plasmids

We obtained *E. coli* K-12 substr. MG1655 *mutS*::FRT from Ivan Matic (Université Paris Descartes, INSERM U1001, Paris, France) through Jesús Blázquez (Centro Nacional de Biotecnología, CSIC, Madrid, Spain) [266]. In this *E. coli* strain, the gene encoding the

protein MutS has been deleted. This protein is a component of the mismatch repair system that recognizes and binds mispaired nucleotides so that the mispairing can be corrected by two further repair proteins, MutL and MutH. The strain MG1655 Δ*mutS* has a predicted mutation rate that is 1000-fold higher than the wild type [419], which ensures that a sufficient number of mutations occur during the mutation accumulation experiment. We transformed this strain with the plasmid pKJE7 (Takara, Cat. #3340), which contains an operon encoding DnaK, and its co-chaperones DnaJ and GrpE under the regulation of a single promoter inducible by L-arabinose [417]. We generated a control strain by transforming the same Δ*mutS* strain with a plasmid that lacks the operon *dnaK–dnaJ–grpE* but is otherwise identical to pKJE7. We refer to this plasmid as pKJE7-DEL(*dnaK–dnaJ–grpE*). This control plasmid was derived from the plasmid pKJE7 by removal of the operon *dnaK–dnaJ–grpE* with a restriction digest using *BamHI* and *SpeI*, followed by religation, after obtaining permission for plasmid modification from Takara.

4.4.2 Evolution experiment

We evolved 68 clonal lines of the hypermutable *E. coli* Δ*mutS* strain containing pKJE7 ($DnaK^+$ lines) and 16 lines containing the control plasmid pKJE7-DEL(*dnaK–dnaJ–grpE*) ($DnaK^-$ lines) by daily passaging them through single-cell bottlenecks on solid LB medium (agar plates; Pronadisa #1551 and #1800) supplemented with 20 mg/ml of chloramphenicol (Sigma- Aldrich #C0378) (fig. 4.1). Except for 8 $DnaK^+$ lines and 8 $DnaK^-$ lines, all the remaining lines were evolved in the presence of 0.2% (w/v) of L-arabinose (Sigma-Aldrich #A3256), which induces the expression of DnaK/DnaJ/GrpE from the plasmid pKJE7 but not from the control plasmid pKJE7-DEL(*dnaK–dnaJ–grpE*). We passaged both the DnaK and $DnaK^+$ lines during 85 days or ~1,870 generations (conservatively assuming ~22 generations per daily growth cycle), except for those lines that went extinct before reaching the end of the experiment. We evolved half of the $DnaK^+$ and $DnaK^-$ lines under mild heat-stress (42 °C) whereas the other half remained at 37 °C.

4.4.3 Verification of DnaK overexpression

We grew the ancestral and evolved strains ($DnaK^+$ and $DnaK^-$, at 37 °C and 42 °C) from glycerol stocks in liquid LB medium supplemented with 20 mg/ml of chloramphenicol in the presence or absence of the inducer L-arabinose (0.2%). After 24 h of growth, we pelleted cells by centrifugation at 12,000 rpm. We resuspended the pelleted cells in 100 ml lysis buffer (containing 200 mM Tris–HCl pH 6.8, 10 mM DTT, 5% SDS, 50% glycerol). To prepare a crude extract, we first boiled resuspended cells at 95 °C for 15 min. After the removal of cell debris by centrifugation, we quantified soluble proteins using the Bradford method [420]. We loaded 1 mg of total protein for each sample in SDS–PAGE gels (12.5% resolving gel). In addition, we loaded onto all gels samples from the ancestral $DnaK^-$ and $DnaK^+$ strains grown in the presence of inducer at 37 °C, as controls to facilitate inter-gel comparisons. We detected DnaK protein by Western blotting using as primary antibody a mouse monoclonal antibody specific to *E. coli* DnaK (Abcam #ab69617) at a 1:10,000 dilution, and as secondary antibody a goat polyclonal (alkaline phosphatase-conjugated) antibody specific to mouse IgG1 (Abcam #ab97237). We scanned membranes after colorimetric detection of conjugated antibodies with the BCIP®/NBT-Purple liquid substrate system (Sigma-Aldrich #BP3679), and used ImageJ to quantify the intensity of DnaK bands on the Western blots [421]. We used the control samples to normalize abundances, which allow the comparison of DnaK levels across experiments.

We examined the change in DnaK levels along a daily cycle of growth for a $DnaK^+$ line evolved at 42 °C (line #2, supplementary fig. 4.2) that showed a lowed DnaK level after ~1,870 generations of mutation accumulation. After 24 h of exponential growth at 42 C in liquid LB medium supplemented with chloramphenicol, we diluted the culture to OD ~0.3, and induced DnaK expression by adding 10 mM of L-arabinose. We allowed the culture to grow for another 24 h in the presence of this expression inducer. Each hour, 1 ml of culture was removed and the DnaK level following the protocol described earlier was measured.

4.4.4 Whole-genome sequencing

We sequenced the genomes of 2 DnaK$^-$ and 6 DnaK$^+$ lines after 85 single-cell bottlenecks. All of these lines evolved in the presence of L-arabinose in the medium, although only DnaK$^+$ cells are able to overexpress DnaK. Half of the sequenced DnaK$^-$ and DnaK$^+$ lines evolved at 37 °C, whereas the other lines evolved at 42 °C. We used the genome sequence of the ancestral Δ*mutS* strain from which both the DnaK$^+$ and DnaK$^-$ lines were derived from our previous study [266].

Specifically, for the evolved lines we performed paired-end Illumina whole-genome sequencing. For DNA extraction, we used the QIAmp DNA mini kit (Qiagen, Venlo [Pays Bas], Germany) in a QiaCube automatic DNA extractor using bacterial pellets obtained from ~10 ml cultures. We constructed multiplexed DNAseq libraries from each clonal evolution line using the TrueSeq DNA polymerase chain reaction-free HT sample preparation kit (Illumina). We performed paired-end sequencing on an Illumina HiSeq2000 platform, using a 2 × 100bp cycles configuration.

We converted sequencing reads from Illumina quality scores into Sanger quality scores. Subsequently, we used the breseq v 0.24rc4 (version 4) pipeline [422] for aligning the Illumina reads to our *E. coli* parental genome and for identifying single nucleotide polymorphisms and indels using bowtie2 [423]. We performed individual runs of breseq, with junction prediction disabled but otherwise default parameters for each of the evolved lines. We deposited the data from this project at the NCBI Sequence Read Archive under the accession SRP074414.

4.4.5 Sequence data

We obtained the complete genomes of *E. coli* K-12 MG1655 (NC_000913) and *S. enterica* serovar Typhimurium LT2 (NC_003197) from GenBank Genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). We also used a data set from Williams and Fares [279] that consists of 1,092 multiple sequence alignments of conserved orthologous proteins from 85 gamma-proteobacterial genomes.

4.4.6 DnaK dependency

We obtained information about DnaK clients from Calloni *et al.* [240]. This study used quantitative proteomics to identify 674 DnaK interactors or client proteins. For 668 of these proteins, the investigators calculated a relative enrichment factor that indicates the fraction of cellular protein bound to DnaK at 37 °C. We used this measure as a proxy for DnaK dependency. We excluded from our analyses the transposases InsC, InsH and InsL of the insertion sequences IS2, IS5 and IS186, respectively. In the genome of *E. coli* K-12 MG1655 there are 6 copies of *insC*, 11 of *insH* and 3 of *insL*.

4.4.7 GroEL dependency

We obtained information about 253 GroEL clients from Kerner *et al.* [249] and Fujiwara *et al.* [250]. Our set of GroEL clients is the union of the slightly different GroEL interactomes characterized in these two studies. We excluded from our analyses the transposase insH of the insertion sequences IS5 and 3 clients reported by Kerner *et al.* [249], which are encoded on plasmids (SwissProt Accession Numbers: P00810, P29368 and Q9339).

4.4.8 Orthology

We identified 3,159 one-to-one orthologs in *E. coli* and *S. enterica* genomes as reciprocal best hits [424] using the Basic Local Alignment Search Tool (BLAST, i.e., BLASTP with an E-value cut-off of 10^{-10}). We identified 631 and 242 *S. enterica* orthologs to DnaK and GroEL clients, respectively. We aligned each pair of orthologous proteins with the Needleman-Wunsch dynamic programming algorithm, using the Needle program from the EMBOSS package [425]. We translated the resulting alignments into codon-based nucleotide alignments with PAL2NAL [426].

4.4.9 Evolutionary rates

We estimated the rate of nonsynonymous substitutions (d_N) using the program codeml from the package PAML 4.7 (one-ratio model M0) [427]. We calculated protein distances

for the gamma-proteobacterial alignments from Williams and Fares [279], using PROTDIST from the PHYLIP package [428] and the Jones, Taylor and Thornton (JTT) substitution matrix [429]. We calculated an average distance for each cluster of orthologous proteins as the mean of all pairwise distances.

4.4.10 Codon usage bias

We computed the Codon Adaptation Index (CAI) using the program CAI from the EMBOSS package [425]. We calculated Codon Usage Bias (CUB) for each pair of *E. coli–S. enterica* orthologs as the mean of the CAI values for each pair of orthologs. We used CUB as a proxy for gene expression.

4.4.11 Protein-protein interactions

We obtained the number of protein-protein interactions (PPI) for each *E. coli* K-12 protein from Rajagopala *et al.* [430]. The binary interactions considered for this study are a combination of the following: (1) literature curated interactions supported by multiple studies or methods and (2) interactions identified by yeast two-hybrid (Y2H) screening. We removed interactions involving DnaK or DnaJ.

4.4.12 Essentiality

We obtained data about gene dispensability for *E. coli* K-12 in rich media from Baba *et al.* [431].

4.4.13 Gene expression

We obtained gene expression data for *E. coli* K-12 MG1655 grown in rich media (LB) at 37 °C from Chen and Zhang [432], where gene expression levels are measured as number of RNA-seq reads per gene length.

4.4.14 Statistical tests

We carried out all statistical analyses and plotted data with R [433] using the packages "base", "pcor", "ggplot2", "dplyr" and "gridExtra".

4.5 Supplementary results

4.5.1 Partial correlation with gene expression

In this analysis we evaluated gene expression as a possible confounding factor of the positive association between DnaK dependency and evolutionary rate. For 457 client proteins for which we have gene expression data, we could carry out a partial correlation analysis to control for the effect of expression without having to resort to a proxy such as CUB. Gene expression level, measured for *E coli* MG1655 in rich media at 37 °C, correlates with CUB ($\rho = 0.321$, $N = 2601$, $P < 2.2 \times 10^{-16}$). A partial correlation analysis shows that the effect of DnaK dependency (measured as a relative enrichment factor on DnaK) on nonsynonymous divergence is independent of gene expression ($\rho = 0.220$, $N = 457$, $P = 1.5 \times 10^{-6}$). The same is observed for protein distances ($\rho = 0.318$, $N = 205$, $P = 1.9 \times 10^{-6}$).

4.5.2 Multiple linear regression for the association between DnaK dependency and evolutionary rates

Because CUB, the number of protein-protein interactions (PPI) and essentiality are important determinants of evolutionary rates, we wished to account for these factors when studying the association between the evolutionary rates (d_N and protein distances) of DnaK clients and their chaperone dependence. To this end, we performed two different multiple regression analyses. Each linear regression had d_N or protein distance as a response variable that depended on five explanatory variables: one categorical (essential = 1, nonessential = 0) and three continuous (PPI, CUB and relative enrichment factor on DnaK). We performed a base-10 logarithmic transformation of variables when such transformations lead to a higher coefficient of determination (R^2). We transformed logarithmically all the continuous variables, except PPI and protein distance. We added a small constant value of 0.001 to d_N, which on occasions was zero, as previously recommended by other authors [434, 435]. The results of these analyses are shown in Supplementary

Table 4.2. Our analysis reproduces the well-known negative correlations between evolutionary rate and both essentiality and CUB [227, 385]. Importantly, they show that DnaK clients more dependent on the chaperone evolve faster with independence of CUB, PPI and essentiality.

4.5.3 DnaK clients evolve slower than nonclients, but strong clients evolve faster than weak clients

If we consider all proteins not included in the set of 674 DnaK clients studied by Calloni *et al.* [240] as nonclients of DnaK, then we find that nonclient proteins evolve faster than clients, both on intermediate time scales (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$; Supplementary Fig. 4.3A), and on long time scales (Wilcoxon rank-sum test, $P = 5.2 \times 10^{-6}$; Supplementary Fig. 4.3B). To evaluate if this difference was caused by a confounding factor we carried out two separate analyses of covariance (ANCOVA). Each ANCOVA had one response variable, d_N or protein distance, and four explanatory variables: two categorical variables (essential = 1/nonessential = 0 and client = 1/nonclient = 0) and two continuous variables (PPI and CUB). For the d_N model we logarithmically transformed CUB and PPI and for the protein distance model we only transformed PPI. The results of these analyses are shown in Supplementary Table 4.3. They indicate that DnaK clients evolve slower than putative nonclients even after controlling for confounding factors. However, there are at least three possible explanations for this result. First, the set of proteins classified as nonclients may contain clients not detected by Calloni *et al.* [240]. Second, the identity of a protein as a client or as a nonclient may have changed drastically during evolution. Third, DnaK clients are intrinsically more constrained than nonclients.

In a subsequent analysis, we subdivided 627 DnaK-interacting proteins in our dataset into strongly interacting clients, those with a relative enrichment factor equal or above the 75% percentile ($N = 158$), and weakly interacting clients ($N = 160$), those with a relative enrichment factor equal or below the 25% percentile. If DnaK buffers destabilizing

mutations, we would expect strong clients to evolve more rapidly. This is indeed the case, both on intermediate time scales (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$; Fig. 4.5A) and on long time scales (Wilcoxon rank-sum test, $P = 2.3 \times 10^{-3}$; Fig. 4.5B). These observations parallel the greater number of mutations that accumulated in strong clients compared to weak clients during the evolution experiment.

Next, we compared strong DnaK clients with weakly interacting clients using two additional and separate ANCOVAs. Each ANCOVA had d_N or protein distance as response variable and the same explanatory variables as before. For the ANCOVA model with d_N as response variable we logarithmically transformed CUB and PPI, and for the model of the ANCOVA with protein distance as response variable we only transformed CUB. The results of these analyses are shown in Supplementary Table 4.4. They indicate that strong clients have a higher nonsynonymous substitution rate than weak clients after controlling for confounding factors. Nevertheless, we do not observe the same pattern in the divergence over long evolutionary time scales, measured as protein distances. In conclusion, controlling for potentially confounding factors such as CUB, essentiality and PPI, strong clients evolve faster on intermediate time scales but not on long time scales, likely due to a lack of statistical power. Also, the quantities in this analysis (CUB, essentiality, etc) are based on *E. coli* and are probably not conserved across the 85 gamma-proteobacterial species we used to calculate protein distances.

4.6 Supplementary tables

Supplementary Table 4.1: Number of synonymous, transitions, transversions, nonsense nucleotide substitutions, and indels accumulated per clonal lineage after ~1,870 generations of evolution in mutation accumulation experiments conducted at 37 °C and 42 °C

Temperature	**Line**[a]	**Number of different types of mutations**				
		Synonymous	**Transitionsk**[b]	**Transversions**[c]	Nonsense	Indels
37 °C	DnaK$^+$ #1	27	104	8	1	20
	DnaK$^+$ #2	68	192	7	6	22
	DnaK$^+$ #3	56	191	11	1	7
	DnaK$^-$	76	214	6	0	28
42 °C	DnaK$^+$ #1	75	214	14	1	40
	DnaK$^+$ #2	77	232	12	0	20
	DnaK$^+$ #3	94	260	13	1	23
	DnaK$^-$	80	224	11	0	32

[a] Experimental evolution lines sequenced in this study. For each temperature, we sequenced three lines overexpressing the DnaK–DnaJ–GrpE chaperone system (DnaK$^+$ lines) and a control line where this system is expressed at wild type levels (DnaK$^-$ line).

[b] To compute the number of transitions and transversions we included single nucleotide substitutions in intergenic regions.

[c] Nonsense mutations were counted as nonsynonymous mutations in Table 4.1.

Supplementary Table 4.2: Main effects in the multiple linear regressions evaluating the influence of different factors on evolutionary rate.

Evolutionary rates	**Factor**[a]	**Slope**[b]
d_N	Protein-protein interactions	-4.643×10^{-3}
	Essentiality	-6.588×10^{-1} (***)
	Codon usage bias	-4.993×10^{-1} (***)
	Relative enrichment factor on DnaK	1.227×10^{-1} (***)
Protein distance	Protein-protein interactions	-2.504×10^{-4}
	Essentiality	-3.858×10^{-2}
	Codon usage bias	-6.076×10^{-1} (***)
	Relative enrichment factor on DnaK	1.440×10^{-1} (**)

[a] The factor in bold is the main focus of this analysis.
[b] Significance levels: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.0001$.

Supplementary Table 4.3: Main effects in the ANCOVAs evaluating the influence of different factors on evolutionary rate.

Evolutionary rates	**Factor**[a]	**Slope**[b]
d_N	Protein-protein interactions	-3.465×10^{-3}
	Essentiality	-3.037×10^{-2} (***)
	Codon usage bias	-7.517×10^{-1} (***)
	DnaK clients / nonclients	-2.372×10^{-2} (***)
Protein distance	Protein-protein interactions	-9.432×10^{-3}
	Essentiality	$-3.739 \times^1 0-2$ (*)
	Codon usage bias	-8.220×10^{-1} (***)
	DnaK clients / nonclients	-4.822×10^{-2} (**)

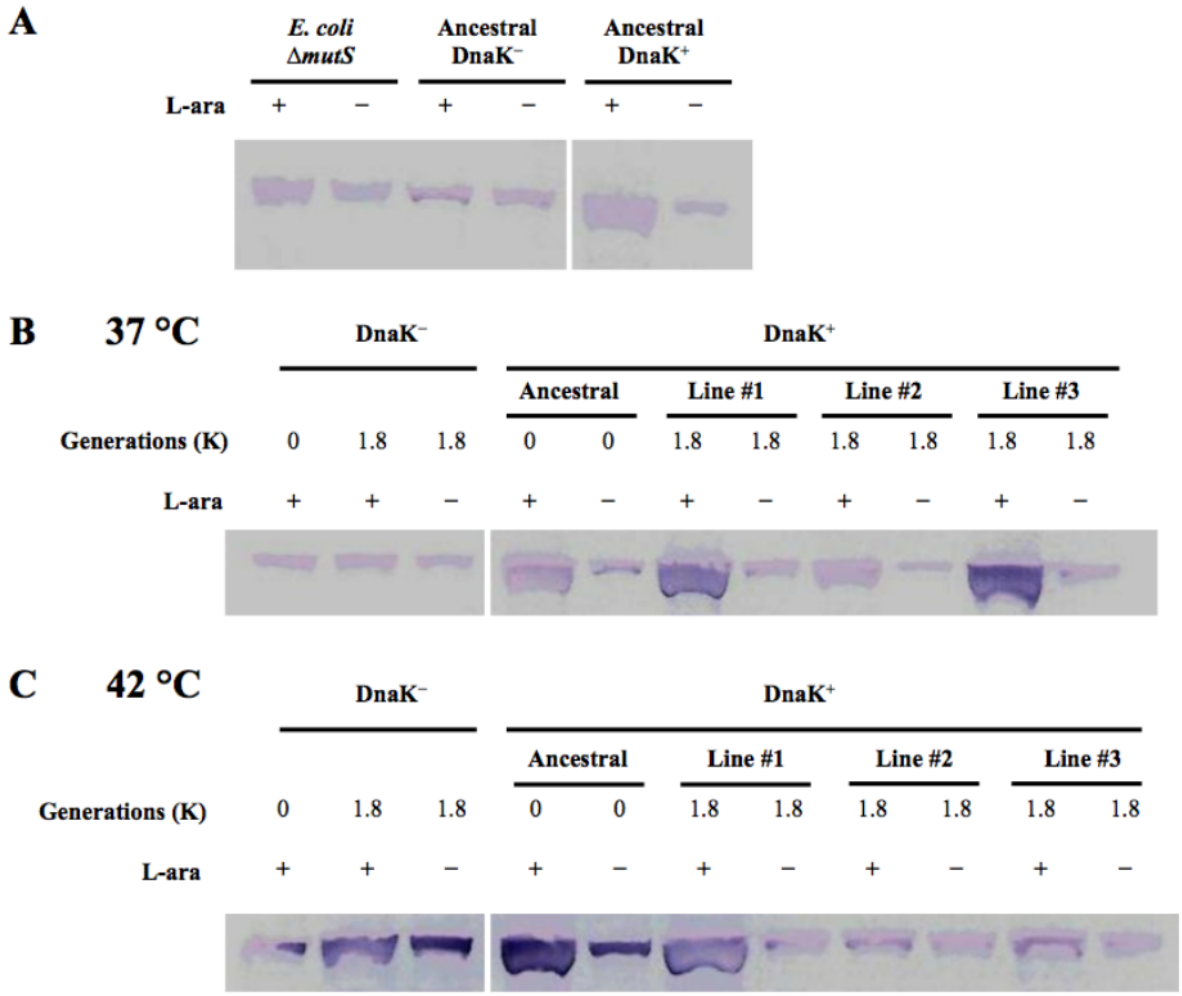
[a] The factor in bold is the main focus of this analysis.
[b] Significance levels: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.0001$.

Supplementary Table 4.4: Main effects in the ANCOVAs evaluating the influence of different factors on evolutionary rate.

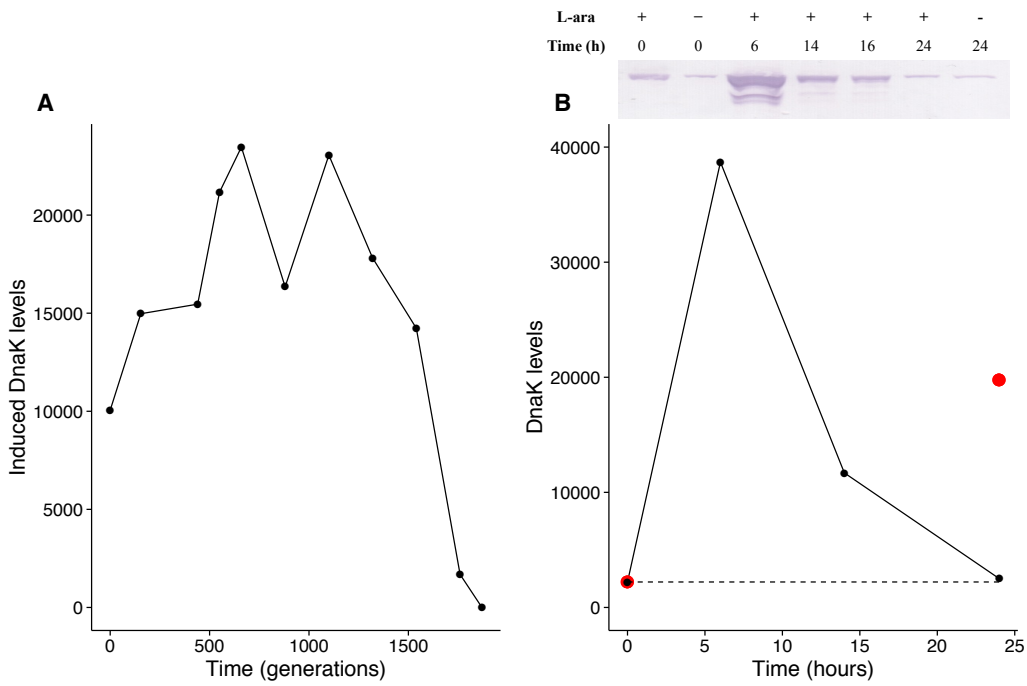
Evolutionary rates	**Factor**[a]	**Slope**[b]
d_N	Protein-protein interactions	-4.222×10^{-2}
	Essentiality	-3.516×10^{-3} (*)
	Codon usage bias	-3.383×10^{-1} (***)
	Strong clients / weak clients	2.164×10^{-2} (**)
Protein distance	Protein-protein interactions	-8.211×10^{-1}
	Essentiality	-1.310×10^{2} (*)
	Codon usage bias	-1.740×10^{3} (**)
	Strong clients / weak clients	8.215×10^{1}

[a] The factor in bold is the main focus of this analysis.
[b] Significance levels: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.0001$.

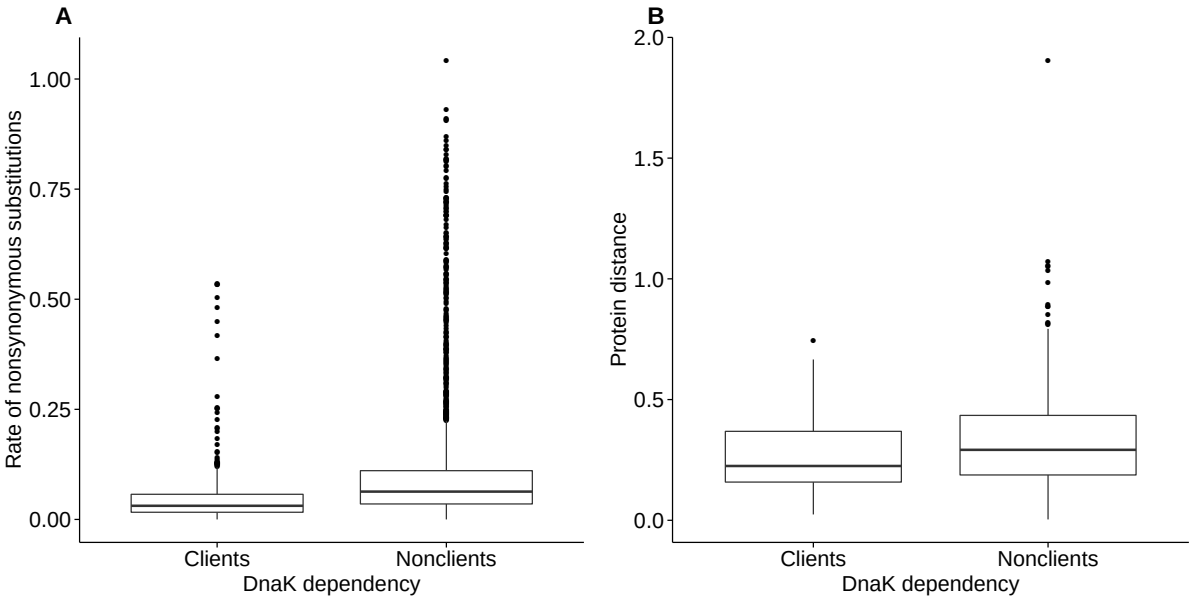
4.7 Supplementary figures



Supplementary Figure 4.1: DnaK levels in Western blots. Representative Western blots for both ancestral and evolved DnaK$^+$ and DnaK$^-$ lines. The strains were grown in the presence (+) or absence (-) of L-arabinose (L-ara), which induces the expression of DnaK from the plasmid pKJE7. Only the DnaK$^+$ strains that harbor this plasmid are able to overexpress the chaperone in the presence of L-arabinose. (**A**) Representative Western blot for the parental hypermutable *E. coli* strain Δ*mutS* and the ancestral DnaK$^+$ and DnaK$^-$ lines. (**B**) Representative Western blot for bacterial lines evolved at 37 °C in a mutation accumulation experiment trough 85 single-cell bottleneck or approximately 1,870 bacterial generations (1.8K). (**C**) Representative Western blot for bacterial lines evolved at 42 °C in a mutation accumulation experiment through the same number of passages, together with ancestral strains growth at the same temperature.



Supplementary Figure 4.2: DnaK levels in the DnaK$^+$ line #2 evolved at 42 °C. (**A**) The evolution of DnaK expression in the DnaK+ line #2 at 42 °C along the mutation accumulation experiment. The induced levels were measured as the intensity of the DnaK band in the Western blot in the presence of the inducer L-arabinose minus the intensity of the band in the absence of inducer. We determined such levels from cultures growth from glycerol stocks stored after the following daily passages: 0, 7, 20, 25, 30, 40, 50, 60, 70, 80 and 85. We obtained the cell extracts from liquid cultures grown for 24h. It is noticeable that the overexpression of DnaK in this line is maintained until ~1760 generations of evolution (daily passage 80), after which the chaperone level decreases to a level found in wild-type *E. coli* and DnaK$^-$ lines. (**B**) The panel demonstrates that not even the loss of DnaK overexpression in line #2 is complete. At passage 85 (~1,870 generations) this line can yield an increment in DnaK abundance similar to the ancestral DnaK$^+$ strain (or the same line at passage 70) during the first 12-16h of growth in liquid culture. Complete loss of overexpression occurs at some point between 16 and 24h of growth. The uppermost part of the panel shows a fragment of the Western membrane used to quantify DnaK abundance over the time course of a daily growth cycle, where we indicate the hour at which we took a sample and the presence (+) or absence (-) of the inducer (L-ara) in the growth medium. We observe that the abundance of DnaK increases until reaching a maximum at 6h after the induction, and then decreases until reaching a level similar to the one found in uninduced DnaK$^+$ lines (dashed line). This level is one order of magnitude lower than the level found in other evolved DnaK$^+$ lines after 24h of growth in liquid culture (red circles indicate the DnaK level of line #3 at 0h and at 24h of growth).



Supplementary Figure 4.3: DnaK clients evolve slower than nonclients. (**A**) We find that nonclients evolve faster than DnaK clients on intermediate evolutionary time scales, measured as the rate of nonsynonymous substitutions (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$). (**B**) On long evolutionary time scales, we also find that nonclients evolve faster than clients (Wilcoxon rank-sum test, $P < 5.2 \times 10^{-6}$). The thick horizontal line in the middle of each box represents the median of the data, while the bottom and top of each box represent the 25^{th} and 75^{th} percentiles, respectively.

5 Chaperonin overproduction and metabolic erosion caused by mutation accumulation in *Escherichia coli*

Submitted as:

Aguilar-Rodríguez J, Fares, M. A., and Wagner A (2017)

Abstract

Bacterial cells adapting to a constant environment tend to accumulate mutations in portions of their genome that are not maintained by selection. This process has been observed in bacteria evolving under strong genetic drift, and especially in bacterial endosymbionts of insects. Here, we study this process in hypermutable *Escherichia coli* populations evolved through 250 single-cell bottlenecks on solid rich medium in a mutation accumulation experiment that emulates the evolution of bacterial endosymbionts. Using phenotype microarrays monitoring metabolic activity in 95 environments distinguished by their carbon sources, we observe how mutation accumulation has decreased the ability of cells to metabolize most carbon sources. We study if the chaperonin GroEL, which is naturally overproduced in bacterial endosymbionts, can ameliorate the process of metabolic erosion, because of its known ability to buffer destabilizing mutations in metabolic enzymes. Our results indicate that GroEL can slow down the negative phenotypic consequences of genome decay in some environments.

5.1 Introduction

Organisms adapting to a constant environment for many generations tend to lose fitness in other environments, because they do not benefit from maintaining fitness in environments they no longer encounter [436–439]. This evolutionary process is known as ecological specialization, and it can be either driven by natural selection or by genetic drift [440–442]. Specialization can be driven by natural selection if an adaptive phenotype shows antagonistic pleiotropy—a phenotype that is selected for being optimal in a given environment is deleterious in other environments due to physiological trade-offs. Alternatively, mutations that are neutral in a particular selective environment can be deleterious in other environments, and such mutations can accumulate in a genome, especially in "unused" genomic regions, by random genetic drift. These two causes of ecological specialization are not mutually exclusive, and sometimes the distinction between them is not clear, for example when neutral "specializing" mutations hitchhike to fixation with beneficial mutations [441].

Specialization through neutral mutations is especially important in bacterial endosymbionts of some insects such as aphids [252–254]. These bacteria live inside host cells that provide a highly constant environment, and have been evolving clonally for millions of generations in this environment. Endosymbionts produce nutrients missing from the host diet, while benefiting from the nutrient-rich and stable intracellular environment of the host. They are maternally inherited by the host, and experience severe bottlenecks during their vertical transmission from generation to generation, which result in small effective population sizes and strong genetic drift. Consequently, natural selection is considerably less efficient in these bacteria than in free-living bacteria. Additionally, a lack of functional DNA repair enzymes (which results in high mutation rates) further accelerate the accumulation of mutations in their genomes, and a lack of recombination prevents the effective purging of deleterious mutations [257].

As a consequence of their high mutational load, such endosymbionts have evolved a mechanism to buffer deleterious mutations. Specifically, they overproduce proteins that

help fold other proteins, molecular chaperones such as GroEL and DnaK, which can reduce the fitness cost of destabilizing mutations in proteins [254, 258–265]. In the rich chemical environment that their host cells provide, selection on many metabolic genes is relaxed in these endosymbionts [252, 253, 255, 443]. Therefore, these genes tend to accumulate loss-of-function mutations that cause metabolic erosion, a process by which an organism loses metabolic abilities, such as the ability to metabolize nutrients.

The *Escherichia coli* Long-Term Evolution Experiment (LTEE) conducted by Richard Lenski and colleagues has offered a great opportunity to study metabolic specialization in a free-living bacterium—the closest free-living relative to the well-studied endosymbiont of aphids *Buchnera aphidicola* [444]. In this experiment, investigators propagated 12 populations of *E. coli* by serial (1:100) dilution for more than 60,000 generations (and counting) in a minimal medium containing glucose as the only source of carbon and energy. The effective population size of these evolving populations is $\sim 3.3 \times 10^7$ [445], which is larger than in bacterial endosymbionts. Adaptation to this single environment has resulted in decreased performance in other environments [440, 442]. Some populations have evolved hypermutable (mutator) phenotypes [446]. In these populations the higher mutation rate caused substantial metabolic erosion, measured as the number of different chemical environments where fitness has declined. This suggests that rather than physiological trade-offs underlying antagonistic pleiotropy, the neutral process of mutation accumulation suffices to cause metabolic specialization in the LTEE [442]. Additionally, the fitness reduction of mutator populations is ameliorated by growth at lower temperature. Most amino acid substitutions are slightly deleterious because of their destabilizing effects, and a mutation's heat sensitivity typically indicates that the mutation affects protein stability [274]. Therefore, temperature-dependent fitness reduction suggests that in these populations metabolic erosion is caused by destabilizing mutations affecting metabolic enzymes [442]. If this is the case, the overproduction of chaperones that can buffer mutations affecting protein stability could have a similar effect to low temperature in ameliorating metabolic erosion caused by mutation accumulation.

GroEL is a member of the family of chaperones known as chaperonins, which are large double-ring complexes that enclose target proteins for folding within a cylindrical folding chamber [214]. GroEL can buffer destabilizing mutations in its target proteins [258, 266, 267, 278–280, 409–411]. Many of these proteins are metabolic enzymes [447]. Enzymes evolved under GroEL overproduction accumulate twice as many mutations, and these mutations have higher destabilizing effects than in the absence of GroEL overproduction [278]. These observations make it plausible that chaperones such as GroEL could slow down the process of metabolic erosion in endosymbionts.

One of the first demonstrations that GroEL can buffer deleterious mutations was observed in *E. coli* populations with high mutational loads [258]. These populations had evolved for more than 3,000 generations via single-cell bottlenecks on solid glucose-limiting minimal medium. Half of the populations evolved at a high mutation rate. The evolution of these populations was dominated by genetic drift, due to the extreme bottlenecks they experienced during their evolution, which are similar to the bottlenecks experienced by endosymbionts during their vertical transmission between hosts. At the end of the experiment, the evolved populations had considerably reduced their fitness (growth rate) in the evolution environment. Moreover, overproducing GroEL in the evolved populations restored fitness to almost ancestral levels. This restoration was observed after supplementing the minimal growth medium with amino acids, otherwise, the energetic cost of overproducing GroEL was so great that no fitness recovery was observed. Similar results were later obtained in mutator populations of *Salmonella typhimurium* [267]. Both studies focused on a single environment. In this study, we investigate whether GroEL overproduction can ameliorate the metabolic erosion experienced by bacteria evolved under conditions that favor the accumulation of mutations in a large number of environments. To address this question, we study mutator *E. coli* populations evolved for thousands of generations under conditions that emulate those of bacterial endosymbionts evolution and we assay their ability to metabolize different carbon substrates using phenotype microarrays.

5.2 Results

Our experiment begins with three independent clonal populations from a mutation accumulation experiment that had been initiated from the same mutator clone (*E. coli* K12 MG1655 Δ*mutS*). The populations had been evolved by daily passaging through single-cell bottlenecks on solid rich growth (LB) medium at 37 °C [448, 449] for 250 days or approximately 5,500 generations (~22 generations between bottlenecks). In such an evolution experiment, the efficiency of natural selection is severely reduced because of the extreme bottlenecks to which the evolving lineages are exposed. In consequence, non-lethal mutations can accumulate freely under the influence of genetic drift [44].

We transformed the ancestor and a clone from each of the three evolved populations with the plasmid pGro7 [417]. This plasmid contains the operon *groE*, which encodes GroEL, and its co-chaperone GroES, under the regulation of a single promoter inducible by L-arabinose. We also transformed the ancestor and each of the three evolved clones with a control plasmid pGro7-DEL(*groE*) (pGro7c) that lacks the operon *groE* but is otherwise identical to pGro7 (Materials and Methods). In the presence of the expression inducer L-arabinose, only cells harboring the plasmid pGro7 overproduce GroEL. In total, we thus generated eight strains from the three evolved clones (E1, E2, and E3) and their ancestor (A). Four of these strains contain the plasmid pGro7 (A/pGro7, E1/pGro7, E2/pGro7, and E3/pGro7), and the other four contain the control plasmid pGro7c (A/pGro7c, E1/pGro7c, E2/pGro7c, and E3/pGro7c). L-arabinose cannot be used as a carbon source by any of these strains.

To study metabolic erosion in the evolved populations we used Biolog phenotype microarrays, which allow the high-throughput measurement of growth and cellular respiration in multiple environments contained in different wells of a 96-well microtiter plate [450, 451]. These arrays measure metabolic activity using tetrazolium, a redox dye that absorbs the electrons from the electron transport chain, and changes color when beign reduced by respiring cells. Biolog assays are widely used, and have been useful to map the phenotypes of various genotypes (e.g., gene knock-out mutants) [450, 452, 453], to

study metabolic innovation [454], to discover new metabolic pathways [455], to reveal an uncoupling between genetic diversity and metabolic phenotypic diversity across *E. coli* strains [456], to study macroevolutionary patterns of phenotypic evolution in bacteria [457], and to characterize the phenotypes of experimentally evolved lineages [458, 459]. Specifically, we used PM1 plates, which measure cellular respiration and growth in 95 chemical environments distinguished by their carbon source, to determine the metabolic phenotypes for each of the eight strains used in this study (Materials and Methods).

We define metabolic erosion as the fraction of carbon sources that an evolved population metabolizes more slowly than the ancestor strain from which the mutation accumulation experiment started. To measure metabolic erosion we only consider carbon sources that the ancestor was able to metabolize above a minimum threshold that is defined by measurement noise (Materials and Methods). We consider that a metabolic phenotype has not declined if it is above this noise threshold, and if it is not significantly lower than the ancestral phenotype (t-test, false discovery rate < 0.05, Materials and Methods). To evaluate the extent of this erosion in the evolved populations, we first focused on the strains that harbor the control plasmid pGro7c, that is, they do not overproduce GroEL. After having experiences 250 single-cell bottlenecks, that is, ~5,500 generations of mutation accumulation, the evolved populations (E1/pGro7c, E2/pGro7c, and E3/pGro7c) fare considerably worse than their ancestor (A/pGro7c) in a majority of the carbon sources tested (between 100% to 90.9%; Table 5.1).

The evolved lineages E1 and E2 had been sequenced in a previous study [448]. Analyzing the sequence data, we find that metabolic erosion and the number of accumulated nonsynonymous mutations are not associated in a straightforward pattern. For example, population E1 accumulated 303 mutations and metabolizes 98.7% of carbon sources more poorly than the ancestor, while population E2 accumulated more (731) mutations and shows metabolic erosion on fewer (90.9%) carbon sources. Of the 304 and 731 nonsynonymous mutations accumulated in lineages E1 and E2, 139 and 277 mutations (45.9% and 37.9%, respectively) affected metabolic enzymes [460]. Lineages E1 and E2 had 182

Table 5.1: Metabolic erosion in the presence and absence of GroEL overproduction for each of three populations evolved in a mutation accumulation experiment for more than 5,000 generations.

Evolved population	**Plasmid**[a]	**Metabolic erosion**[b]
E1	pGro7c	98.7% (76/77)
E1	pGro7	88.6% (62/70)
E2	pGro7c	90.9% (70/77)
E2	pGro7	75.7% (53/70)
E3	pGro7c	100% (77/77)
E3	pGro7	95.7% (67/70)

[a] Cells harboring the plasmid pGro7 overproduce GroEL in the presence of the expression inducer L-arabinose, while cells harboring the control plasmid pGro7c do not.
[b] The fraction of those carbon sources on which evolution caused significant metabolic erosion, based on carbon sources which the ancestor metabolizes above the noise threshold of 0.039 (Materials and Methods).

and 224 metabolic pathways (43% and 53%, respectively) affected by nonsynonymous mutations [460]. The observation that E2 cells express GroEL and DnaK more highly than E1 cells may help explain the poor correlation between metabolic erosion and the number of accumulated mutations [449]. Similarly to GroEL, DnaK is a chaperone that can help buffer the effects of mutation, and that is highly expressed in bacterial endosymbionts [281, 282].

To find out whether chaperones can indeed mitigate metabolic erosion, we turned to the mutation accumulation populations that overproduce GroEL. To analyze data from these populations, however, we needed to take into account that chaperone overproduction carries energetic costs [258, 266]. The ATP consumed in the reaction cycle of the chaperonin contributes less to this cost than the synthesis of large amounts of GroEL and GroES proteins [216, 258]. The cost is evident by comparing the ancestral strains A/pGro7 with A/pGro7c, where GroEL overproduction causes a decline in metabolic rates in most environments (Supplementary Fig. 5.1). To take this cost into account, we compared the metabolic phenotypes of evolved strains overproducing GroEL (E1/pGro7, E2/pGro7, and E3/pGro7) with the corresponding ancestor (A/pGro7). We find that

populations overproducing GroEL show less metabolic erosion than populations with no GroEL overproduction (Table 5.1). However, this reduction in metabolic erosion is small (it occurs in only 9.9% of environments on average). We observe that populations with a higher level of metabolic erosion show lower amelioration of this erosion after GroEL overproduction (Table 5.1). However, this correlation is not significant (Pearson's correlation coefficient = -0.910, $P = 0.273$), probably because it is based on little data (n = 3).

5.3 Discussion

Bacteria evolving for many generations in the same chemical environment experience a process of metabolic specialization by which they lose or reduce their ability to metabolize many different metabolic substrates. The evolution of bacteria with low effective population sizes is dominated by genetic drift. In such populations, drift drives the fixation of deleterious mutations, reducing a population's ability to grow in many environments where the ancestor was able to grow. Arguably, bacterial endosymbionts of some insect species provide the best well-known examples of this process [255]. These intracellular bacteria specialize in synthetizing essential nutrients missing from the host diet. They have such low effective population sizes and high mutation rates that they experience an irreversible process of genome decay and reduction, by which they mutate and finally lose segments of their genomes that are not maintained by selection. This process of genome erosion is so relentless that endosymbionts can even lose functions essential for their symbiotic role. Eventually, a second endosymbiont can join such a symbiotic consortium, which makes the genes shared by both redundant. In such cases, the endosymbiont that has lost some metabolic pathway which is essential for the symbiosis, may become replaced by the "healthier" endosymbiont [252, 255]. In some cases where no secondary endosymbiont exists, the decline of metabolic phenotypes can cause substantial fitness reductions in the host [255].

Mutation accumulation experiments mimic some of the conditions present during the evolution of intracellular bacteria. In such evolution experiments, bacteria evolve through

single-cell bottlenecks that reduce the efficiency of natural selection. Bacteria evolved in this type of experiment acquire some characteristics common to endosymbionts [449], such as high mutational loads [266, 281, 461], genome reduction [462], or higher expression of molecular chaperones such as GroEL and DnaK [267, 449]. In this study, we analyzed mutator populations subject to daily single-cell bottlenecks for more than 5,000 generations in a nutrient-rich environment. At the end of this mutation accumulation experiment, these populations had also evolved some similarities with bacterial endosymbionts. They showed similar patterns of genome reduction, mutational biases, and gene expression [449].

Previous studies of bacteria evolved in mutation accumulation experiments have shown that they can experience great fitness reductions, which can be partially compensated by the overproduction of chaperones [258, 266, 267, 281]. However, these studies did not explore fitness reduction in environments different from the environment in which evolution took place. Here, using phenotype microarrays, we assayed evolved populations in 95 different environments distinguished by their carbon sources. We find that bacterial populations evolved under strong genetic drift can lose or reduce their ability to metabolize many distinct carbon sources. Because these populations have evolved under conditions that considerably reduce the power of natural selection to drive the fixation of beneficial mutations, the most likely cause for metabolic erosion is the accumulation of mutations driven by genetic drift, rather than the fixation of advantageous mutations that show antagonistic pleitropy [442].

However, we note that there is no correlation between the number of accumulated nonsynonymous mutations and metabolic erosion in two of the evolved populations for which we have whole-genome sequences from a previous study [448]. There is also no correlation between the number of nonsynonymous mutations in metabolic enzymes, and metabolic erosion. There are at least two possible explanations for why the relationship between mutational load and erosion may not be straightforward. First, neutral or slightly deleterious mutations in metabolic enzymes can be highly pleiotropic because of the organization of metabolism as a reaction network. A loss-of-function mutation in a highly connected

enzyme can simultaneously and negatively affect several distinct metabolic phenotypes. The second non-exclusive possibility is that the relationship between mutational load and erosion may be influenced by chaperone expression. We find that the population with a higher chaperone expression shows less metabolic erosion [449]. This observation is in agreement with previous observations that chaperones can partially restore fitness of bacteria evolved in mutation accumulation experiments [258, 266, 267, 281]. The effect of chaperone expression on metabolic erosion may be mediated through the buffering of destabilizing mutations in metabolic enzymes [278]. Such mutations may also cause the erosion observed in populations from the LTEE [442].

GroEL and the Hsp70 chaperone DnaK are the most abundant proteins in the cytosol of endosymbiotic bacteria [254, 259–265]. For instance, GroEL expression is 7.5 times higher in *B. aphidicola* than in its close free-living relative *E. coli* [259]. These observations suggest that GroEL and DnaK help these cells cope with their high mutational loads by buffering the negative effects of destabilizing mutations in proteins, and this hypothesis has experimental support [258, 266, 267, 281]. Here, we find some further evidence that chaperones could ameliorate some of the negative effects of metabolic erosion. In particular, we observe that evolved populations overproducing the chaperonin GroEL show less metabolic erosion than in the absence of chaperonin overproduction. However, we only observe this mitigating effect of GroEL overproduction in some environments. This is not remarkable if we consider that GroEL is not an all-powerful molecular machine, capable of buffering any possible deleterious mutation. The buffering ability of GroEL is most certainly restricted to certain proteins and types of mutations. However, it would be worth exploring if GroEL overproduction could show greater amelioration of metabolic erosion in populations evolved for fewer generations, and therefore with smaller mutational loads.

In summary, we show that bacterial populations evolving in conditions emulating the evolution of bacterial endosymbionts experience a severe reduction in their ability to metabolize many distinct carbon substrates. We find that the relationship between this metabolic erosion and mutational load is not straightforward, and could be influenced by

both pleiotropy and chaperone expression. Finally, we show that GroEL overproduction can mitigate metabolic erosion in some environments. In so doing, we provide evidence that molecular chaperones can reduce in more than one environment the negative impact of genome decay on cell metabolism.

5.4 Materials and Methods

5.4.1 Strains and plasmids

Strains E1, E2, and E3 derive from *E. coli* K12 substr. MG1655 Δ*mutS*. They evolved in parallel in a long-term mutation accumulation experiment on solid LB medium at 37 °C [448, 449], where each population was passaged for 250 days through a single-cell (clonal) bottleneck after 24 hours of growth.

We derived the control plasmid pGro7c from the plasmid pGro7 (Takara, Cat. #3340) via PCR amplification using primers TGATAACTCTCCTTTGAGAAAGTCCG and TTGCCCTGCACCTCGCAGAAATAA, and Phusion® High-Fidelity DNA polymerase (NEB), after having obtained permission from Takara to modify pGro7. We digested the PCR products with Dpn1 to remove pGro7. We separated the digested PCR products on a 0.8% agarose gel, excised the band corresponding in size to the desired amplification product, purified the product using the QIAquick Gel extraction kit (50928704, Quiagen), and quantified it using a Nanodrop spectrophotometer. We generated pGro7c by ligating the amplification product using T4 DNA ligase (NEB). We validated the sequence of pGro7c using Sanger sequencing. We transformed the ancestor (A) strain, and the evolved lines (E1, E2, E3) with pGro7 and pGro7c to generate eight different strains: A/pGro7, E1/pGro7, E2/pGro7, E3/pGro7, A/pGro7c, E1/pGro7c, E2/pGro7c, and E3/pGro7c.

5.4.2 Phenotype microarrays

We used Biolog phenotype microarray PM1 (Biolog, Inc., Hayward, California, USA) to assay the carbon utilization phenotypes of each of the eight strains on 95 different carbon sources. Both cell growth and respiration contribute to these metabolic phenotypes,

because respiration can be independent of growth [442]. We performed four replicate Biolog assays for each strain, for a total of $8 \times 4 = 32$ microarrays. To do so, we streaked each strain from glycerol stocks onto LB agar plates supplemented with 25 μg/mL of chloramphenicol (Sigma-Aldrich #C0378). We incubated the plates at 37 °C for 24 hours, and re-streaked the colonies onto fresh plates of the same type, which we incubated at the same temperature for the same period of time. We resuspended the colonies from the latter plates in IF-0 solution (Biolog, Inc., Hayward, California, USA) using sterile cotton swabs, and then centrifuged the suspension at 3,000 g for 3 min. We resuspended the cell pellet in fresh IF-0 to an optical density at 600 nm (OD_{600}) of approximately 0.05 (as measured in a 200 μL suspension volume). We diluted (1:5) this suspension in IF-0+dye (Biolog, Inc., Hayward, California, USA) supplemented with 25 μg/mL of chloramphenicol, L-methionine (5 μM), cyanocobalamine (0.125 μM), and 0.2% (w/v) L-arabinose (Sigma-Aldrich #3256). We added 100 μL of the resulting solution to each well of a PM1 microarray, which we incubated without shaking at 37 °C for 24 hours. We measured OD_{600} and optical density at 750 nm (OD_{750}) at 0 min and 24 hours using a microplate reader (Tecan Spark 10M). We performed 9 reads per well in a 3-by-3 square grid, and computed the average.

Both the presence of oxidized tetrazolium and that of cells cause absorbance at 600 nm. The absorbance at 600 nm due to the presence of cells can be removed by subtracting the OD_{750} from the OD_{600}, because tetrazolium has almost no absorbance at 750 nm. We therefore used these cell-density corrected values in our study ($OD_{600-750}$), and computed the metabolic phenotype M_S for substrate S as $OD_{600-750,24h} - OD_{600-750,0h}$, where $OD_{600-750,24h}$ is the corrected optical density after 24 hours, and $OD_{600-750,0h}$ is the corrected optical density at the start of the experiment. We used a minimum threshold to detect respiration in a given substrate S. To obtain this threshold we computed the absolute differences between all pairs of 3,072 $OD_{600-750,0h}$ values (8 strains $\times$ 4 replicates $\times$ 96 wells). Because at time 0, cells have not started to metabolize yet, these differences must be caused by experimental noise. We only consider differences in M_S as significant

if they are greater than 0.039, which is the 98-th percentile of all the differences between wells with no growth.

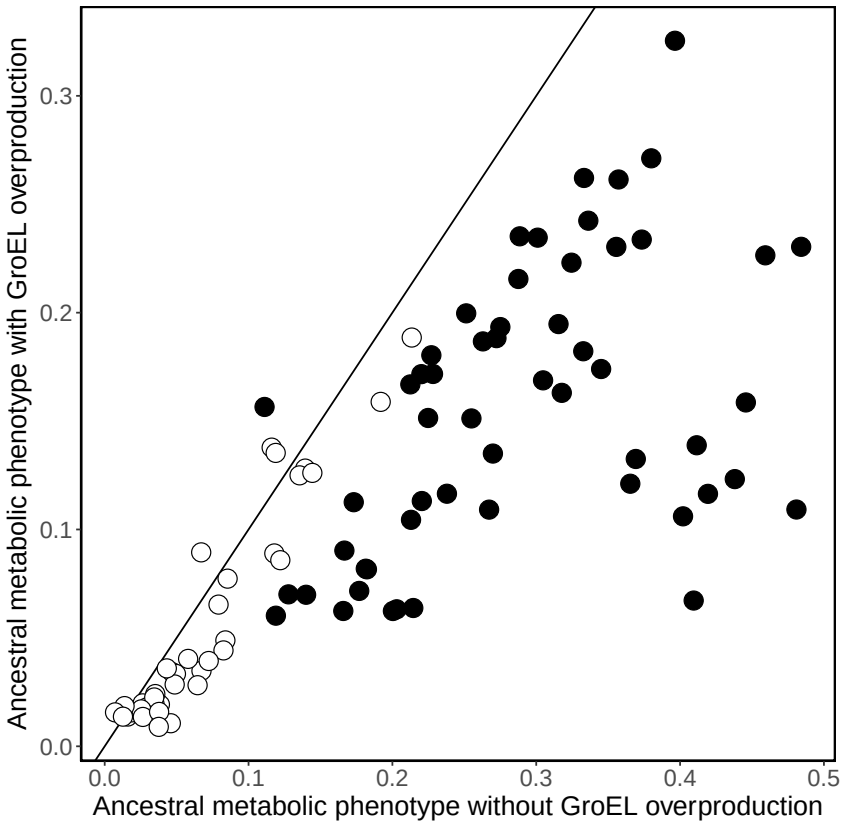
For each evolved strain we compared the four experimental measurements of M_S with the corresponding measurements for the ancestor using a two-tailed t-test. In particular, we compared the three evolved populations with pGro7 (E1/pGro7, E2/pGro7, and E3/pGro7) against A/pGro7, and the three evolved populations with pGro7c (E1/pGro7c, E2/pGro7c, and E3/pGro7c) against A/pGro7c. We adjusted the P values for multiple testing using the false discovery rate method (FDR). We consider that a metabolic phenotype has not declined significantly if $P > 0.05$, and both the evolved and the ancestral average M_S values are greater than 0.039 (the noise threshold).

We quantified metabolic erosion in a given evolved population as

$$\frac{P - C}{P}. \tag{5.1}$$

P refers to the number of carbon sources where M_S is greater than the noise threshold in A/pGro7 for evolved strains overexpressing GroEL, and in A/pGro7c for evolved strains that do not overproduce GroEL. C refers to the number of carbon sources where the metabolic phenotype has not declined significantly in the evolved population.

5.5 Supplementary figures



Supplementary Figure 5.1: GroEL overproduction incurs metabolic costs. . Metabolic phenotypes of the ancestor on 95 different carbon sources. Each circle represents the rate at which the ancestor can metabolize one of the carbon sources averaged over four technical replicates of the ancestor without GroEL overproduction (horizontal axis) and with GroEL overproduction (vertical axes). The diagonal line indicates equal metabolic rates with and without GroEL overproduction. Circles below the diagonal line represent carbon sources where GroEL overproduction causes a reduction in metabolic rate. Filled circles indicate environments where the difference in metabolic rate with and without chaperonin overproduction is statistically significant, i.e., greater than experimental measurement noise.

6 Metabolic determinants of enzyme evolution in a genome-scale bacterial metabolic network

Submitted as:

Aguilar-Rodríguez J, and Wagner A (2017)

Abstract

Different genes and proteins evolve at very different rates. To identify the factors that explain these differences is an important aspect of research in molecular evolution. One such factor is the role a protein plays in a large molecular network. Here, we analyze the evolutionary rates of enzyme-coding genes in the genome-scale metabolic network of *Escherichia coli* to find the evolutionary constraints imposed by the structure and function of this complex metabolic system. Central and highly connected enzymes appear to evolve more slowly than less connected enzymes, but we find that they do so as a by-product of their high abundance, and not because of their position in the metabolic network. In contrast, enzymes catalyzing reactions with high metabolic flux—high substrate to product conversion rates—evolve slowly even after we account for their abundance. Moreover, enzymes catalyzing reactions that are difficult to by-pass through alternative pathways, such that they are essential in many different genetic backgrounds, also evolve more slowly. Our analyses show that an enzyme's role in the function of a metabolic network affects its evolution more than its place in the network's structure. They highlight the value of a system-level perspective for studies of molecular evolution.

6.1 Introduction

Different proteins evolve at very different rates [119, 463, 464]. Half a century after this observation seeded the field of molecular evolution [119]), the reasons are still a subject of active research, and even more so since the genome-era made sequence and functional data about proteins abundantly available. Much of the variation in evolutionary rates stems from variation in selective constraints on proteins, and several factors influence these constraints (for recent reviews, see [464] and [385]). The most important is the amount of a protein that is expressed, and the breadth of its expression across cells or tissues in multicellular organisms [222, 224, 465]. Highly and broadly expressed genes are under strong purifying selection, and therefore evolve slowly. Other factors influence evolutionary rates more weakly. They include protein length [390, 466–469], essentiality [227, 470, 471], multifunctionality [472–475], subcellular localization [476], or being a chaperone client [279–282]. To gain deeper insights into the determinants of protein evolution, one must go beyond a gene-centered approach and embrace a systems-oriented view of protein evolution.

Inside a cell, proteins often form large and complex networks of interacting molecules. The position of a protein within such a network, as well as its role in the network's function, can affect the protein's evolution. In other words, the structure and function of a molecular network can impose selective constraints on its member proteins [477]. For example, proteins at the center of a protein-protein interaction network evolve more slowly (they are more constrained) than those at the periphery [478–483]. In contrast, in the yeast transcriptional regulation network, more central transcription factors evolve faster than less central ones [484]. As these two types of cellular networks have similar topological properties [485], this difference in selective constraints over the network structure must ultimately be caused by different network functions. Nonetheless, despite being significant and consistent across many different organisms, the effects of network topology on protein evolution is weak, and could be caused by confounding factors such as expression level, and it can be affected by biased and low-quality data [486, 487].

Metabolic networks constitute another important class of cellular network. They are well-studied in model organisms such as *Escherichia coli* [285], and comprise hundreds to thousands of chemical reactions, most of them catalyzed by enzymes encoded in genes. In a metabolic network, chemical reactions are organized in a highly reticulate manner to perform two main functions: Energy production and biosynthesis. Specifically, using energy and chemical elements from environmental nutrients, metabolic networks synthetize essential small molecules (i.e., amino acids, ribonucleotides, deoxynucleotides, lipids, and enzyme cofactors). The chemical reactions a metabolic network catalyzes are encoded in a metabolic genotype—a genome's set of enzyme-encoding genes. The network's phenotype can be defined as the set of molecules it can synthesize, and the rate at which it does so [62]. Thanks to computational approaches such as flux balance analysis (FBA) [296, 300], the relationship between metabolic genotypes and phenotypes can be studied computationally, which also allows us to study how selection for a given metabolic phenotype can constrain metabolic enzyme evolution. This type of analysis is currently not possible in other types of molecular networks, such as protein-protein interaction networks.

Previous work in eukaryotes has revealed that more central and more highly connected enzymes in metabolic networks, that is, those sharing metabolites with many other enzymes, evolve more slowly [308, 488–491]. Additionally, enzymes catalyzing reactions with a high metabolic flux—the rate at which a reaction transforms substrates into products—tend to evolve slowly [308, 492]. In the present study, we study how the structure and function of a bacterial metabolic network affects the evolution of metabolic genes through point mutations. To our knowledge, this is the first time that such a study is performed using the whole-genome metabolic reconstruction of *E. coli* [285], which is arguably the best known metabolic network of any living organism. Specifically, we study how quantities such as enzyme connectivity and metabolic flux affect evolutionary rate. To do so, we account for possible flux variation with Markov chain Monte Carlo (MCMC) sampling, a method that has not been used before in this type of evolutionary analysis. Additionally, we also study for the first time the influence of factors such as reaction su-

peressentiality [493], which quantifies how easily a reaction can be bypassed in a metabolic network by other reactions or pathways, and the number of different chemical reactions that an enzyme catalyzes (enzyme multifunctionality). In performing these analyses, we comprehensively characterize metabolic determinants of enzyme evolution in *E. coli*.

6.2 Results

6.2.1 The effect of metabolic network topology on enzyme evolution

To study how network structure affects enzyme evolution, we constructed a *reaction graph* representation of the whole-genome *E. coli* metabolic network, in which the nodes represent reactions. Two reactions are connected by an edge if they share at least one metabolite (Material and Methods). In such a graph, the connectivity of a reaction corresponds to the number of other reactions that produce or consume the reaction's substrates or products. The connectivity of an enzyme is equivalent to the connectivity of the reaction catalyzed by the enzyme. The centrality of an enzyme can be measured as the number of shortest pathways passing through the reaction node associated with the enzyme (betweenness centrality).

In a metabolic network, highly connected enzymes tend to occupy a central position in the network (as determined by their betweenness centrality, Material and Methods), while less connected enzymes are more peripheral (Fig. 6.1A; Spearman's $\rho = 0.519$, $P < 2.2 \times 10^{-16}$, $n = 659$). In other words, enzymes in central metabolic processes, such as central carbon metabolism, tend to be highly connected, while enzymes in peripheral pathways tend to be less connected.

One might expect that more highly connected enzymes in a metabolic network are more constrained in their rate of evolution than less connected enzymes. The reason is that the reaction products of highly connected enzymes are substrates of many different reactions, such that any mutation disturbing product formation is bound to be more deleterious in a highly connected enzyme. However, a previous study on *E. coli* metabolism

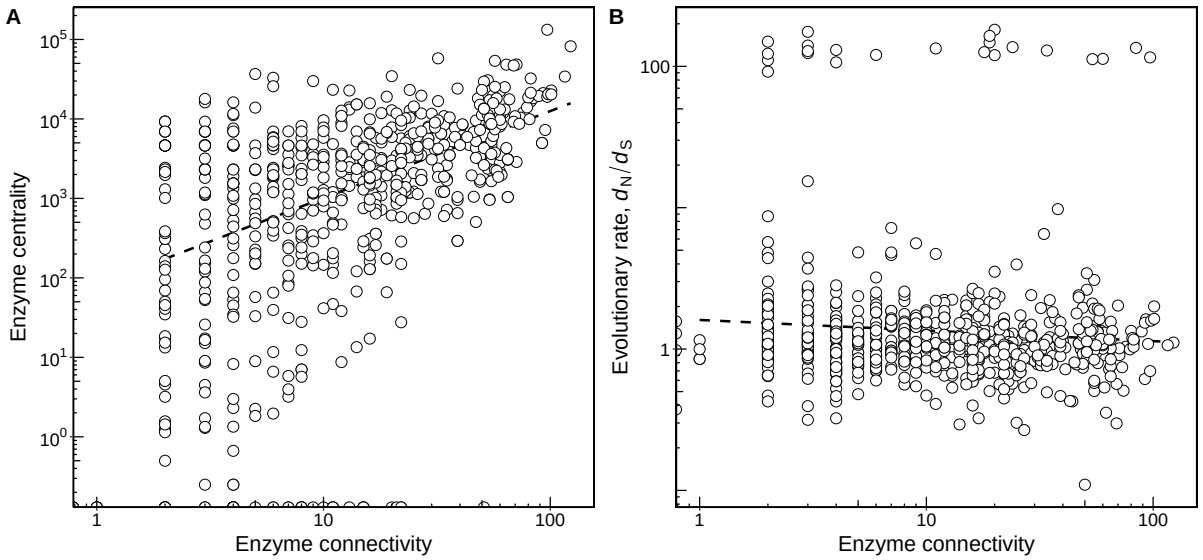


Figure 6.1: Highly central and connected enzymes in a metabolic network do not evolve slowly. (**A**) The relationship between enzyme connectivity and centrality in the *E. coli* metabolic network (Spearman's $\rho = 0.519$, $P < 2.2 \times 10^{-16}$, $n = 659$). The centrality measure of a reaction is its betweenness centrality determined from the reaction graph (Materials and Methods). (**B**) The relationship between enzyme connectivity and evolutionary rate measured as d_N/d_S (Spearman's ρ = -0.094, $P = 0.016$, $n = 659$). In both panels, a dashed line shows the best linear fit to the data and is provided as a visual guide. Note the double-logarithmic scale.

found no correlation between enzyme connectivity in core intermediary metabolism and evolutionary rate, determined as the rate of amino acid replacements, for 108 pairs of *E. coli – Haemophilus influenzae* orthologs [494]. In contrast, a later study found that highly connected enzymes in the metabolic network of *Saccharomyces cerevisiae* do evolve more slowly [308]. We suspected that the original negative result in *E. coli* could be caused by small statistical power resulting from the many fewer enzymes analyzed by Hahn *et al.* [494] ($n = 108$) than by Vitkup *et al.* [308] ($n = 671$). We therefore repeated the *E. coli* analysis using the much larger whole-genome metabolic reconstruction. We estimated the evolutionary rate of an enzyme as the ratio of nonsynonymous substitutions to synonymous substitutions per nucleotide site (d_N/d_S) in the gene coding for the enzyme. We used values of d_N/d_S obtained by comparing genes in *E. coli* to orthologs in the closely related genome of *Salmonella enterica*. A small value of d_N/d_S indicates a lower evolutionary rate due to higher constraints on enzyme evolution. Figure 6.1B shows the relationship between enzyme connectivity and the rate of evolution (Spearman's ρ = -0.094, $P =$

0.016, $n = 659$; Table 6.1). The negative correlation is very small but significant.

Table 6.1: Correlations of various quantities with d_N/d_S.

Quantity	**Spearman's ρ**
Betweenness centrality	0.080^{*}
Enzyme connectivity	-0.094^{*}
Metabolic flux	-0.287^{***}
SI_{glu}	-0.341^{***}
SI_{54}	-0.297^{***}
Gene expression	-0.351^{***}
Protein abundance	-0.501^{***}

Significance levels: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 10^{-}6$.

One potentially important confounding factor in the association between enzyme connectivity and evolutionary constraint is enzyme expression. Highly connected enzymes tend to be highly abundant (Spearman's $\rho = 0.155$, $P = 1.8 \times 10^{-4}$, $n = 581$), and in general, abundant proteins tend to evolve more slowly [222, 224]. This association between expression level and evolutionary rate also holds for enzymes. Specifically, we observe that high enzyme expression is associated with slow evolution (low d_N/d_S) regardless of whether expression is measured on the mRNA level (Spearman's ρ = -0.351, $P = 3.6 \times 10^{-16}$, $n = 508$; Table 6.1) or on the protein level (Spearman's ρ = -0.501, $P < 2.2 \times 10^{-16}$, $n = 581$; Table 6.1). Since expression of enzyme-coding genes is strongly correlated between the mRNA and protein level (Spearman's $\rho = 0.434$, $P < 2.2 \times 10^{-16}$, $n = 444$), we focus our analysis below on the protein level [495], but note that all reported results also hold for the mRNA level. When controlling for enzyme abundance in a partial correlation analysis between enzyme connectivity and evolutionary rate, the correlation loses statistical significance (Spearman's $\rho = 0.007$, $P = 0.865$, $n = 444$; Table 6.2). In other words, while highly connected enzymes evolve at slightly lower rates than less connected enzymes, this association is a byproduct of the relationship between evolutionary rate and enzyme abundance.

Similarly to enzyme connectivity, one might expect that more central enzymes should be more constrained in their evolution, but this relationship is also not consistent across

Table 6.2: Partial correlations of various quantities with d_N/d_S.

Quantity | Controlled quantity	**Spearman's ρ**
Enzyme connectivity | Protein abundance	0.007
Betweenness centrality | Protein abundance	0.080
Metabolic flux | Protein abundance	-0.178^{**}
Metabolic flux | SI_{glu}	-0.172^{**}
SI_{glu} | Protein abundance	-0.212^{***}
SI_{glu} | Metabolic flux	-0.220^{***}
SI_{54} | Protein abundance	-0.195^{**}
SI_{54} | Metabolic flux	-0.209^{***}

Significance levels: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 10^{-}6$.

studies. Some studies in eukaryotic species have found a significant association [488, 490], while others have not [489, 491]. We find a very weak positive association that is barely significant (Spearman's $\rho = 0.080$, $P = 0.040$, $n = 508$; Table 6.1), and that also loses its significance after controlling for enzyme abundance in a partial correlation analysis (Spearman's $\rho = 0.080$, $P = 0.061$, $n = 444$; Table 6.2). Thus, the association between enzyme centrality and evolutionary rate also stems from the relationship between evolutionary rate and enzyme abundance.

6.2.2 Enzymes catalyzing reactions with high metabolic flux evolve slowly

A reaction's metabolic flux refers to the rate at which the reaction converts substrates into products. One might expect that enzymes catalyzing high flux reactions may evolve more slowly. The reason is that such enzymes tend to supply products to a large number of reactions and pathways, such that the effects of flux-diminishing mutations may be more deleterious than in low-flux enzymes [308]. To study the relationship between metabolic flux and the rate of enzyme evolution, we applied flux balance analysis (FBA) to the metabolism of *E. coli* [285]. FBA is a linear programming method that maximizes the rate of biomass production in a given nutritional environment, simultaneously balancing all the metabolic fluxes under a steady state assumption and a set of flux constraints [296]. FBA has been extensively used to predict the phenotype of a metabolism from its

genotype, that is, to predict the ability of a metabolism to synthetize biomass in a given chemical environment from the genes encoding the metabolism's enzymes [62, 92, 286, 290, 300, 457, 496, 497]. FBA predictions are in good agreement with experimental data for model organisms such as *E. coli* [285, 498–503].

We applied FBA to the *E. coli* metabolic network iAF1260 [285], maximizing aerobic growth on glucose in an environment where glucose is the only carbon source. Analyzing the association between metabolic flux and evolutionary rate is complicated by the fact many distributions of fluxes through individual enzymes can produce the same maximal biomass synthesis rate. For example, if two different reactions can produce the same biomass molecule at the same maximal rate, one of the two reactions could carry the maximal flux, while the other carries no flux, or both reactions could be active, such that the sum of their individual fluxes produces the metabolite at the maximal rate. In other words, a metabolic network can solve the problem of synthesizing biomass in multiple equivalent ways. To account for this flux variation, we used MCMC sampling to uniformly sample the space of all possible flux values [302]. We then computed a distribution of flux values for each of the reactions in the *E. coli* metabolic network, and used the median of this distribution as the reaction flux. To our knowledge, this is the first time that the complete flux distribution, as determined by MCMC sampling, is taken into consideration in studying the relationship between metabolic flux and enzyme evolution.

Figure 6.2 shows that enzymes catalyzing high-flux reactions evolve more slowly (Spearman's $\rho =$ -0.323, $P < 2.2 \times 10^{-16}$, $n = 614$; Table 6.1). Importantly, this association does not disappear if we account for enzyme abundance: While high-flux enzymes tend to be highly abundant (Spearman's $\rho =$ -0.375, $P < 2.2 \times 10^{-16}$, $n = 543$), they still evolve more slowly in a partial correlation analysis that controls for enzyme abundance (Spearman's $\rho =$ -0.178, $P = 3 \times 10^{-5}$, $n = 543$; Table 6.2). This observation agrees with a previous finding that high-flux yeast metabolic enzymes are subject to more constrained evolution [308]. A similar association has been found with experimental flux measurements in the human erythrocyte core metabolism [492].

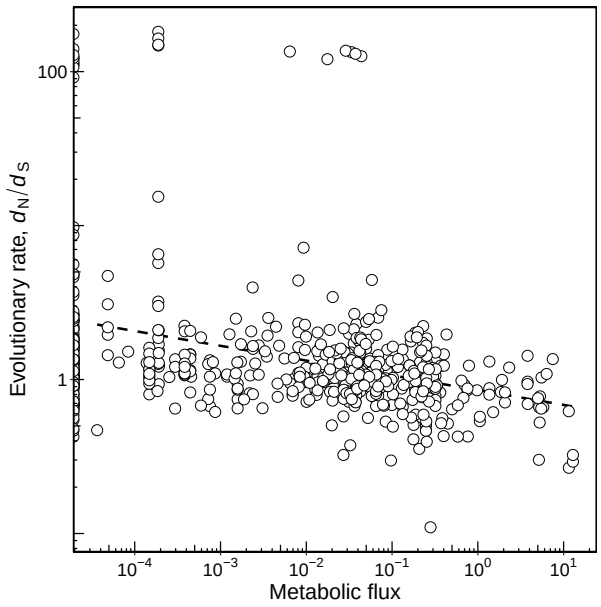


Figure 6.2: Enzymes catalyzing reactions with high metabolic flux evolve more slowly. The relationship between metabolic flux and enzyme evolutionary rate measured as d_N/d_S (Spearman's ρ = -0.323, $P < 2.2 \times 10^{-16}$, $n = 614$). The dashed line shows the best linear fit to the data and is provided as a visual guide. Note the logarithmic scale on both axes.

6.2.3 Highly superessential enzymes evolve slowly

A central function of a metabolic network is to synthetize the small-molecule precursors of biomass (amino acids, nucleotides, cofactors, etc.) that are indispensable for cell growth and survival. In a given chemical environment, a metabolic reaction is *essential* if its product is needed for viability, i.e., for biomass synthesis, and if its removal ("knock-out") eliminates this ability. Otherwise the reaction is nonessential. Reaction essentiality depends not only on the environment, but also on a network's genotype, that is, on the genes encoding the enzymes of the network. For example, certain genes are only essential in some strains of *Saccharomyces cerevisiae* [504]. One reason for such variation in essentiality is that different organisms can synthesize the same biomass molecules via alternative metabolic pathways that comprise different biochemical reactions and enzymes, which are encoded by different genes [290, 498, 505, 506].

While it is easy to manipulate an organism's environment experimentally to study how reaction essentiality depends on the environment, current technologies limit our ability to

systematically alter metabolic genotypes to study how essentiality varies with metabolic genotypes, i.e., with the presence or absence of genes encoding alternative metabolic pathways. This limitation calls for computational approaches. One such approach is suited to study comprehensively how the presence or absence of enzyme-coding genes affects the essentiality of other enzyme-coding genes [290]. It builds on the ability of FBA to efficiently predict a metabolic network's phenotype—whether the network can produce biomass in a given environment—from its genotype. Briefly, the approach samples the "universe" of more than 5,000 biochemical reactions known to occur in at least one species, to generate viable metabolic networks with a given phenotype, but an otherwise random complement of reactions [62]. By analyzing large ensembles of such random viable networks, one can determine how difficult it is to bypass a reaction through an alternative metabolic pathway, by computing a reaction's *superessentiality index* (SI) [290]. The SI of a reaction, which ranges from zero to one, is the fraction of random viable networks in which the reaction is essential for viability. In any given environment, reactions with a SI close to zero are easily bypassed, and non-essential for viability in most metabolisms, whereas reactions with the highest SI of one are always essential and cannot be bypassed according to current biochemical knowledge.

It is possible that highly superessential reactions (large SI, not easily by-passed) evolve at lower rates, because they may be subject to stronger purifying selection caused by their greater importance for viability in different genetic backgrounds. This could be especially the case in bacteria, where gene content can evolve very fast via lateral gene transfer, so that a given enzyme may become part of many different metabolic networks during its evolutionary history. To find out whether this is the case, we used superessentiality indices of *E. coli* metabolic reactions computed for (i) an aerobic minimal environment with glucose as the only carbon source (SI_{glu}) and (ii) 54 minimal environments that contain different unique carbon sources (SI_{54}) [290].

Figure 6.3A shows that *E. coli* reactions with high SI_{glu} evolve more slowly (Spearman's ρ = -0.341, $P < 2.2 \times 10^{-16}$, $n = 568$; Table 6.1). It is possible that this association

could be explained by enzyme abundance, because superessential enzymes tend to be highly abundant (Spearman's $\rho = 0.302$, $P = 1.6 \times 10^{-12}$, $n = 525$). However, the association between SI_{glu} and d_N/d_S persists in a partial correlation analysis that controls for protein abundance (Spearman's $\rho = -0.212$, $P = 9.7 \times 10^{-7}$, $n = 525$; Table 6.2). In other words, enzymes that are difficult to bypass in a glucose minimal environment evolve slowly, and do so independently of their abundance.

Like SI_{glu}, SI_{54} quantifies how difficult it is to bypass a metabolic reaction, but does so for 54 different environments, each containing one of 54 nutrients as its sole carbon source. A reaction or enzyme has a high SI_{54} if its removal abolishes viability in at least one of the 54 different environments for a large fraction of random networks viable in these 54 environments. Enzymes with a high SI_{54} also evolve slowly (Fig. 6.3B; Spearman's $\rho = -0.297$, $P = 5.3 \times 10^{-13}$, $n = 568$; Table 6.1). While these enzymes also tend to be highly abundant (Spearman's $\rho = 0.208$, $P = 1.6 \times 10^{-6}$, $n = 525$), the association persists when we control for enzyme abundance in a partial correlation analysis (Spearman's $\rho = -0.195$, $P = 7.2 \times 10^{-6}$, $n = 525$; Table 6.2).

Reactions highly superessential in a glucose-minimal environment tend to carry a high metabolic flux in this environment (Spearman's $\rho = 0.516$, $P < 2.2 \times 10^{-16}$, $n = 568$). Metabolic flux is thus an additional potentially confounding factor for the observed relationship between SI_{glu} and evolutionary rate. However, a partial correlation analysis shows that enzymes with high SI_{glu} still evolve more slowly after controlling for metabolic flux (Spearman's $\rho = -0.220$, $P = 1.3 \times 10^{-7}$, $n = 568$; Table 6.2). Similarly, the effect of SI_{54} on enzyme evolution still holds after controlling for metabolic flux (Spearman's $\rho = -0.220$, $P = 1.3 \times 10^{-7}$, $n = 568$; Table 6.2).

6.2.4 The multifunctionality of an enzyme does not affect its rate of evolution

Metabolic enzymes can be classified as either specialists or generalists [307]. A specialist enzyme catalyzes one specific chemical reaction, while a generalist enzyme catalyzes more

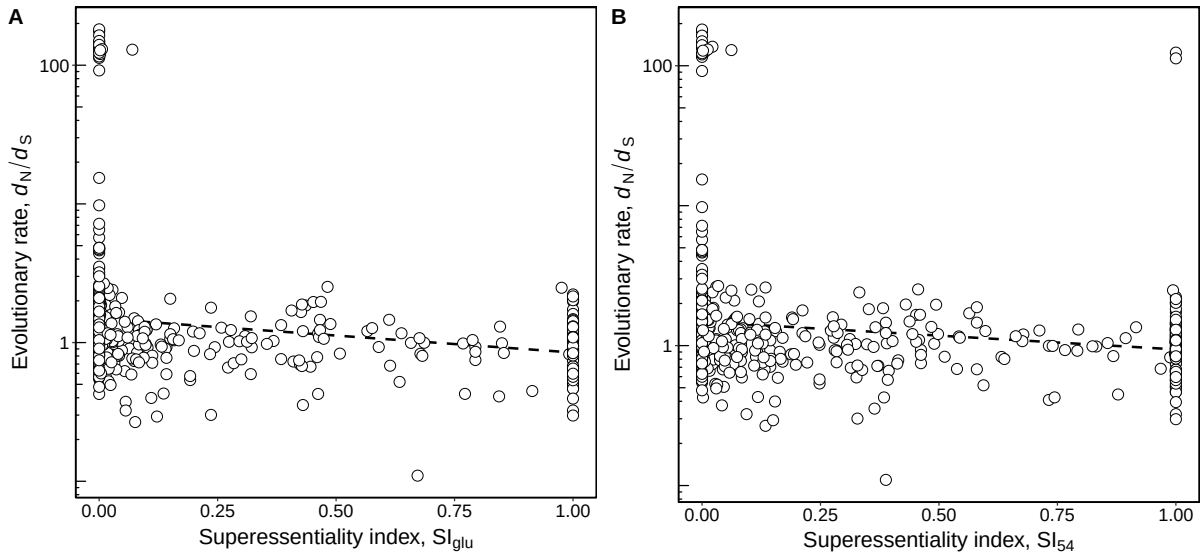


Figure 6.3: Enzymes with high superessentiality evolve more slowly. (**A**) Scatter-plot showing the negative association between enzyme superessentiality in glucose (SI_{glu}) and evolutionary rate measured as d_N/d_S (Spearman's ρ = -0.341, $P < 2.2 \times 10^{-16}$, n = 568). (**B**) Scatter-plot showing the association between enzyme superessentiality in 54 different carbon sources (SI_{54}) and d_N/d_S (Spearman's ρ = -0.297, $P = 5.3 \times 10^{-13}$, n = 568). In both panels, a dashed line shows the best linear fit to the data and is provided as a visual guide. Note the logarithmic scale of the y-axes.

than one reaction. One might expect that generalist enzymes evolve more slowly than specialist enzymes, since mutations in the genes encoding them may affect more than one metabolic pathway or function. This would at least be predicted by existing work on mutations that are pleiotropic, i.e., they affect multiple different phenotypes [507]. For example, theoretical considerations [508–510], and empirical evidence in yeast suggest that highly pleiotropic mutations tend to be more deleterious than less pleiotropic mutations [511].

For metabolic enzymes in *E. coli*, we find that generalist enzymes have a lower median evolutionary rate (1.096; n = 220) than specialist enzymes (1.114; n = 424), but the difference between these two enzyme categories is not significant (Wilcoxon rank-sum test, P = 0.747). Thus, there is no connection between multifunctionality or pleiotropy on the one hand, and evolutionary rate on the other hand, at least for *E. coli* metabolic enzymes.

6.3 Discussion

Natural selection on the function of a molecular network constrains how the network's genes evolve. Conversely, changes in network genes affect the function of the whole network. In other words, the evolution of a network's *parts* affects the evolution of the *whole* network, and vice versa. These two types of influence are entangled, because changes in network function that result from changes in network genes can themselves impose new evolutionary constraints on network genes. Here we study how the structure and function of a large metabolic network (the whole) influences the evolution of its constituent enzymes (the parts). In doing so, we perform a comprehensive exploration of the metabolic determinants of enzyme evolution. Our analysis is part of a research tradition aiming to understand the molecular evolution of living systems by relating the evolutionary rates of genes with their function and position in a biological network [308, 385, 464, 478, 482, 483, 494]. An advantage of using metabolic systems in such studies is that the relationship between the functions of the enzymes and the network is especially well understood [20, 284, 300].

First, we show that the position of an enzyme in the *E. coli* metabolic network does not affect its rate of evolution. Previous studies have found significant but very modest correlations between some topological network parameters and evolutionary rates in other metabolic networks and pathways [308, 488–490]. However, in the *E. coli* metabolic network, central and highly connected enzymes do not evolve at different rates when we control for their abundance. This corroborates previous findings in small-scale metabolic systems of mammals [490, 492] and *E. coli* [494]. Other studies in yeast [308] and *Drosophila* [489] have found that the connectivity of an enzyme influences its rate of evolution. However, even where significant, this association is very weak. Such a weak or absent association is not unreasonable, considering the "bow-tie" architecture of a metabolic network [291, 292], where numerous input pathways of nutrient conversion feed into a highly interconnected central core metabolism, which feeds many output biosynthetic pathways. Some of these biosynthetic pathways are linear sequences of reactions that produce essential and complex biomass molecules, such as amino acids or enzyme

cofactors. A loss-of-function mutation of an enzyme in one such linear and peripheral pathway would be lethal [51], even though the enzyme is not highly connected. In other words, mutations in both central and peripheral enzymes can be deleterious, albeit for different reasons.

In agreement with previous studies in other organisms [308, 492], we find that metabolic flux—the rate at which a reaction converts substrates into products—affects enzyme evolution. We find that enzymes catalyzing reactions with high flux tolerate fewer amino acid substitutions than enzymes catalyzing reactions with lower fluxes. In other words, the function of a metabolic network, that is, biomass production, constrains the evolution of network genes through amino acid substitutions in a non-uniform way: Enzymes with high flux experience greater constraints than enzymes with low flux, since they are more important for network function.

In any one metabolic network, a loss of function mutation in a given enzyme may be lethal (in a specific environment), because it abolishes the network's ability to produce biomass. In other metabolic networks with the same phenotype but a different metabolic genotype—a different complement of enzyme-coding genes—the enzyme may not be essential, because alternative reactions or pathways can assume its role. The extent to which an enzyme or reaction is easy or difficult to bypass is a function of metabolic biochemistry, and can be quantified through a reaction's superessentiality index [290]. Highly superessential reactions (enzymes) are difficult to bypass and their loss would be lethal in many different genetic backgrounds, while the loss of lowly superessential enzymes would be lethal in only a few backgrounds.

We find that highly superessential enzymes evolve more slowly. Relevant for this observation is that the metabolic genotypes of bacteria can evolve very rapidly. That is, bacterial enzymes can rapidly get lost via gene deletion or loss-of-function mutations, and new enzymes may be acquired via horizontal gene transfer [312]. For example, closely related *E. coli* strains may differ in more than 20% of their genomes, and in hundred or more metabolic genes, a difference that is partly due to horizontal gene transfer and

gene deletions [313, 512]. On evolutionary time scales, bacterial metabolic enzymes can thus find themselves operating in different genotypic backgrounds, such that differences in superessentiality matter for their rate of evolution, as our data shows. Superessentiality might influence the rate of evolution less in organisms whose metabolic genotypes change more slowly.

Finally, we also tested if generalist enzymes, which catalyze many reactions, are subjected to higher selective constraints than enzymes just catalyzing a single chemical reaction, as theoretical expectations would predict [508–510]. Previous studies have found that multifunctional genes in yeast evolve slowly [473, 474], corroborating theoretical expectations [513], although the magnitude of this effect is very modest. In mammals, multifunctional proteins also tend to be constrained, and the more functions a protein is involved in, the lower is its rate of evolution [475]. However, generalist (multifunctional) enzymes do not evolve more slowly, indicating that pleiotropy is not constraining enzyme evolution, at least in *E. coli*.

We note that myriad other, non-metabolic factors may influence the evolution of enzyme-coding genes. These include protein structure [514], chaperone targeting [279–282], and many others, but the dominant factor is usually gene expression level [385, 464]. It is thus remarkable that the associations between evolutionary rate and metabolic flux or superessentiality are moderately high, comparable in strength to that between evolutionary rate and mRNA expression level, and only below the association between evolutionary rate and protein abundance.

In conclusion, our analysis of the rates of evolution of enzyme-coding genes in the *E. coli* metabolic network shows how a gene's role in the function of a larger network can affect its evolution. In doing so, we show how a systems-level perspective can help understand the factors that contribute to protein evolution.

6.4 Materials and Methods

6.4.1 Metabolic network

To investigate how the topology of a metabolic network affects the evolution of metabolic genes, we constructed a reaction graph representation of the *E. coli* metabolic network model iAF1260 [285], which includes 2,382 reactions and 1,972 metabolites. In a reaction graph, nodes represent reactions, which are connected by an edge if they share at least one metabolite as either a substrate or a product [515]. When constructing this reaction graph, we did not consider the following currency metabolites, which are the most highly connected metabolites: H, H_2O, ATP, orthophosphate, ADP, pyrophosphate, NAD, NADH, AMP, NADP, NADPH, CO_2, and CoA [308]. The inclusion of such metabolites, which participate in many different reactions, would create many reactions that are adjacent in the graph but not otherwise functionally related. Such reactions would come to dominate the structure of the network, and obscure patterns of connections between functionally related reactions. The reaction graph thus created comprises 2,382 nodes and 18,953 edges. In this graph, we computed the connectivity (or degree) of every reaction, which is its number of edges. In other words, the connectivity of a reaction is the number of other reactions that share at least one metabolite with the focal reaction. To determine the centrality of a reaction, we computed its betweenness centrality [381, 516], which is the number of shortest paths between any two nodes that pass through this reaction.

To study how different properties of a metabolic reaction may affect the evolution of the enzyme-coding gene whose product catalyzes the reaction, it is preferable to work mostly with reactions that show a one-to-one relationship to enzyme-encoding genes. Therefore, we exclude from our evolutionary analyses reactions catalyzed by large macromolecular complexes that are encoded by multiple genes. Following Vitkup *et al.* [308], for enzymes that catalyze more than one reaction, we use the reaction carrying the largest metabolic flux (the rate at which metabolites are converted into products) because it is the reaction imposing a higher evolutionary constraint. In addition, also following Vitkup *et al.* [308],

wherever different enzymes (isoenzymes) catalyze the same chemical reaction, we use the enzyme with the lowest rate of sequence evolution. The resulting dataset comprises 659 enzyme-coding genes associated with the same number of metabolic reactions.

6.4.2 Metabolic fluxes

We determined the distribution of fluxes that is allowable during growth on glucose for each reaction in the *E. coli* metabolic model iAF1260 [285] using MCMC sampling [302]. We used the artificially centered hit-and-run algorithm (ACHR) [303] with minor modification as described by Bordbar *et al.* [304] and Lewis *et al.* [305]. We implemented the ACHR algorithm with the ACHRSampler in COBRA Toolbox v.2.0.5 [517], using the in the MATLAB (The MathWorks, Natick, MA) environment R2012b. We used a minimal (computational) medium in which glucose was the only carbon source, and set the uptake rate of glucose to the value of 8 millimoles per gram dry cell weight per hour. Following Nam *et al.* [307], in order to restrict the sampling to the space of flux values relevant to *in vivo E. coli* growth on glucose, we established a lower bound to the biomass objective function of 90% of the optimal growth rate predicted by FBA [296]. The mixed fraction is a metric introduced by Bordbar *et al.* [304] to measure the uniformity of the sample from the space of allowed fluxes. We obtained a mixed fraction of 0.5096, which suggests that the space was nearly uniformly sampled [304]. We removed reactions with a median flux value greater than 15 millimoles per gram dry cell weight per hour from further analysis to ensure the exclusion of reactions involved in futile cycles [306, 518].

6.4.3 Reaction superessentiality and enzyme multifunctionality

We obtained superessentiality indices of metabolic reactions for growth on glucose (SI_{glu}) and for growth on 54 different sole carbon sources (SI_{54}) from Barve *et al.* [290].

We followed the classification of E. coli K-12 enzymes in generalists and specialists of Nam *et al.* [307]. Enzymes that only catalyze a specific chemical reaction were classified as specialists, while enzymes that catalyze more than one reaction were classified as generalists.

6.4.4 Evolutionary rates

We obtained the values of d_N/d_S, d_N, and d_S in this analysis from the study by Alvarez-Ponce *et al.* [448]. In that study, orthologs in *E. coli* and *S. enterica* genomes were identified as reciprocal best hits [424] using the protein-protein Basic Local Alignment Search Tool (i.e., BLASTP with an *E*-value cut-off of 10^{-10}). Each pair of orthologous proteins was aligned using ProbCons 1.2 [519]. The resulting alignments were back-translated into codon-based nucleotide alignments, and the ratio d_N/d_S was estimated using the program codeml from the package PAML 4.7 (one-ratio model M0) [427].

6.4.5 Gene expression and protein abundance

We obtained gene expression data for *E. coli* K-12 MG1655 grown in rich medium (LB) at 37 °C from Chen and Zhang [432], who quantified gene expression levels as numbers of RNA-seq reads per gene, normalized by gene length. We retrieved protein abundance data of *E. coli* K-12 MG1655 from the integrated dataset of PaxDb 3.0 [495].

Curriculum vitae

Personal Information

Name: José Aguilar-Rodríguez

Date of birth: January 28, 1988

Nationality: Spanish

ORCID ID: 0000-0001-7622-5482

Education

University of Zurich, Switzerland; 2012 - 2017

PhD in Natural Sciences (Evolutionary Biology)

University of Valencia, Spain; 2006 - 2011

"Licenciatura" (5-year degree), Biology

Overall average grade: 9.88/10, Honours system: 3.95/4, *First Class Honours*

I.E.S. Camp de Morvedre, Sagunto, Spain; 2004 - 2006

Baccalaureate degree in Science

Overall average grade: 9.9/10, University entrance exam: 9.56/10

Grants and Awards

- SIB Best Swiss Bioinformatics Graduate Paper Award, 2017 (CHF 5,000)
- Swiss National Science Foundation Early Postdoc Mobility, 2017 (CHF 100,950)
- Helmsley Fellowship, Cold Spring Harbor Laboratory, 2017 ($ 3,000)
- EON Seed Grant (co-applicant), 2016 ($ 50,000)

- Forschungskredit, University of Zurich, 2014 (CHF 57,546)
- National Award for Excellence in Academic Performance in Science, First Prize, Ministry of Education, Spain, 2014 (€3,300)
- "Premio Extraordinario de Licenciatura," University of Valencia, Spain, 2012
- "Carmen y Severo Ochoa" Studentship, Valencia City Council, Spain, 2011 (€12,000) (declined)
- Collaboration Fellowship, Ministry of Education, Spain, 2010-2011 (€1,000)
- Gold medal in the iGEM competition on Synthetic Biology organized by the Massachusetts Institute of Technology, 2010

Publications

11. *__Aguilar-Rodríguez, J.__, *Payne, J. L., Wagner, A. (2017). "A thousand empirical adaptive landscapes and their navigability." *Nature Ecology & Evolution*, 1: 0045.
10. Blattner, A., __Aguilar-Rodríguez, J.__, Kränzlin, M., Wagner, A., Lehner, C. F. (2017). "Drosophila *Nnf1* paralogs are partially redundant for somatic and germline kinetochore function." *Chromosoma*, 126: 145-163.
9. *__Aguilar-Rodríguez, J.__, *Sabater-Muñoz, B., Montagud-Martínez, R., Berlanga, V., Alvarez-Ponce, D., †Wagner, A., †Fares, M. A. (2016). "The molecular chaperone DnaK is a source of mutational robustness." *Genome Biology and Evolution*, 8: 2979-2991.
8. Khalid, F., __Aguilar-Rodríguez, J.__, †Wagner, A., †Payne, J. L. (2016). "Genonets Server – A web server for the construction, analysis, and visualization of genotype networks." *Nucleic Acids Research*, 44: W70-76.
7. Sprouffske, K., __Aguilar-Rodríguez, J.__, Wagner, A. (2016). "How archiving by freezing affects the genome-scale diversity of *Escherichia coli* populations." *Genome Biology and Evolution*, 8: 1290-1298.

6. SIB Swiss Institute of Bioinformatics Members (including **José Aguilar-Rodríguez**) (2016). "The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases." *Nucleic Acids Research*, 44: D27-37.

5. *Sabater-Muñoz, B., *Prats-Escriche, M., Montagud-Martínez, R., López-Cerdán, A., Toft, C., **Aguilar-Rodríguez, J.**, Wagner, A., Fares, M. A. (2015). "Fitness trade-offs determine the role of the molecular chaperonin GroEL in buffering mutations." *Molecular Biology and Evolution*, 32: 2681-2693.

4. Ashikaga, H., **Aguilar-Rodríguez, J.**, Gorsky, S., Lusczek, E., Marquitti, F. M. D., Thompson, B., Wu, D., Garland, J. (2015). "Modelling the heart as a communication system." *Journal of the Royal Society Interface*, 12: 20141201.

3. Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., **Aguilar-Rodríguez, J.**, Vicente-Ripolles, V., Fuster, G., Bernet, G. P., Maumus, F., Munoz-Pomer, A., Sempere, J.M., Latorre, A., Moya, A. (2011). "The *Gypsy* Database (GyDB) of mobile genetic elements: release 2.0." *Nucleic Acids Research*, 39: D70-74.

2. Baquero, F., **Aguilar-Rodríguez, J.**, Moya, A. (2011). "La biología como forma de representación: Wittgenstein y la partitura de la vida." *Ludus Vitalis: Journal of Philosophy of Life Sciences*, 19: 29-46. [In Spanish]

1. The International Aphid Genomics Consortium (including **José Aguilar**) (2010). "Genome sequence of the pea aphid *Acyrthosiphon pisum*." *PLoS Biology*, 8: e1000313.

* Equal contribution

† Co-corresponding authors

Acknowledgments

Darwin's five-year-long voyage around the world aboard H.M.S. Beagle was an extraordinary educational experience that set his scientific career on a successful course. Doing a PhD is the closest experience to Darwin's expedition for most modern scientists. Working in my thesis has been a great adventure, no less full of fascinating challenges, and great moments. Embarking on a voyage of scientific discovery aboard a ship of the Royal Navy is no enterprise which can be undertaken alone. Similarly, completing a PhD is not easily done without fellow explorers. Modern science is an eminently social endeavor, and this thesis owes much to many people.

My deepest level of gratitude goes to three amazing scientists. *All For One and One For All.* First of all, I would like to thank my supervisor Andreas Wagner for his guidance and support, for being a constant source of inspiration, and for all the academic freedom he gave me. The opportunity he granted me to work in his laboratory changed my life, and made me the scientist I now am. Second, I would like to thank Joshua Payne. I feel extremely lucky to have been able to work closely with him. His help has been instrumental for the successful completion of this thesis, and his guidance has been constant in all the project we have undertaken together. He is both a great colleague, and a friend. Finally, I would like to thank Mario Fares for his always constant support. He is a very creative researcher, and a wonderful person. I feel fortunate to have been able to collaborate with him in several fruitful projects. In all of them, I have learnt much from him.

Part of the pleasure of doing science is the great opportunity to interact on a daily basis with extraordinary individuals. Therefore, thanks to all the "Wagnerians," past and present. Special thanks to Aditya Barve, Tugce Bilgin, Sinisa Bratulic, Riddhiman

Dhar, Jordi van Gestel, Josh Payne, Magdalena San Román, Kathleen Sprouffske, and Macarena Toll-Riera. It has been truly great to work in such a fantastic environment. Magdalena is one of the main persons to thank for such a congenial work place, and I thank her for being such a good friend. I feel lucky to have shared an office with Kathleen and Sinisa for a few years. Thanks to Kathleen for teaching me much about experimental evolution. I really enjoyed our collaborations. Thanks to Sinisa for important discussions about science, experiments, and much else besides. I am also grateful to him for helping me around the lab, and for helping me with the formatting of this thesis. Thanks to Macarena for her support, and for her feedback in some of my manuscripts. I learned so much from her, and I thank her for her friendship. I thank Yolanda Schaerli for teaching me the nuts and bolts of working in the wet lab.

The work presented in this thesis would not have been possible without the contributions of my co-authors. My heartfelt thanks to all of them, but special thanks go to David Alvarez-Ponce, Leto Peel, Beatriz Sabater-Muñoz, and Massimo Stella. I would like to acknowledge Martin Ackermann, Erik van Nimwegen, and Owen Petchey for helpful discussions and suggestions on my work. Special thanks go to Martin for sharing his enthusiasm for science in every meeting we had. I owe special thanks to Mar Albà and Sara Mitri for agreeing to review my thesis. I thank Michael Baumgartner for the summary in German. I am also indebted to the Forschungskredit program of the University of Zurich for funding (FK-14-076).

Many thanks to Andrés Moya for being such an incredible mentor, and for all his personal support throughout the years. If he had not recommend me *Robustness and Evolvability in Living Systems* during my second year at college, probably I would never have come to Zurich to pursue my PhD studies with Andreas.

Most importantly, I am incredibly thankful to Marta, for sharing this adventure with me. I would not be here today writing these lines without her constant love, patience, and encouragement. This thesis is dedicated to her, and to my family, who nurtured my curiosity with books: The beginning of it all.

Bibliography

[1] Johannsen, W (1909). *Elemente der Exakten Erblichkeitslehre [The elements of an exact theory of heredity].* Gustav Fischer, Jena.

[2] Johannsen, W (1911). The genotype conception of heredity. *The American Naturalist*, **45**:129–159.

[3] Johannsen, W (1905). *Arvelighedslaerens elementer (The Elements of Heredity).* Gyldendal Boghandel, Copehagen.

[4] Botstein, D (2015). *Decoding the Language of Genetics.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

[5] Schwartz, J (2008). *In Pursuit of the Gene: From Darwin to DNA.* Harvard University Press, Cambridge, MA.

[6] Lewontin, RC (1974). *The genetic basis of evolutionary change.* Columbia University Press, New York.

[7] Provine, W (2001). *The origins of theoretical population genetics.* The Univeristy of Chicago Press, Chicago, edition with a new afterword.

[8] Crow, JF and Kimura, M (1970). *An introduction to population genetics theory.* Harper & Row, New York.

[9] Ewens, W (2004). *Mmathematical Population Genetics. I. Theoretical Introduction.* Springer, New York, 2nd edition.

[10] Gillespie, JH (2004). *Population Genetics: A Concise Guide.* Johns Hopkins University Press, Baltimore, 2nd edition.

[11] Hartl, DL and Clark, AG (2007). *An introduction to population genetics theory*. Sinauer Associates, Sunderland, MA, 4th edition.

[12] Wagner, A (2011). *The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems*. Oxford University Press, New York.

[13] Nei, M (2013). *Mutation-driven evolution*. Oxford University Press, Oxford, UK.

[14] Maynard Smith, J, Burian, R, Kauffman, S, Alberch, P, Campbell, J, Goodwin, B, Lande, R, Raup, D, and Wolpert, L (1985). Developmental Constraints and Evolution: A Perspective from the Mountain Lake Conference on Development and Evolution. *The Quarterly Review of Biology*, **60**(3):265–287.

[15] Gould, SJ (2002). *The Structure of Evolutionary Theory*. Harvard University Press, Cambridge, MA.

[16] Medina, M (2005). Genomes, phylogeny, and evolutionary systems biology. *Proceedings of the National Academy of Sciences*, **102**:6630–6635.

[17] Koonin, EV and Wolf, YI (2006). Evolutionary systems biology: links between gene evolution and function. *Current Opinion in Biotechnology*, **17**:481–487.

[18] Dean, AM and Thornton, JW (2007). Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Reviews Genetics*, **8**:675–688.

[19] Loewe, L (2009). A framework for evolutionary systems biology. *BMC Systems Biology*, **3**:27.

[20] Papp, B, Notebaart, Ra, and Pál, C (2011). Systems-biology approaches for predicting genomic evolution. *Nature Reviews Genetics*, **12**:591–602.

[21] Soyer, O, editor (2012). *Evolutionary Systems Biology*. Springer, New York.

[22] O'Malley, MA (2012). Evolutionary Systems Biology: Historical and Philosophical Perspectives on an Emerging Synthesis. In OS Soyer, editor, *Evolutionary Systems Biology*, Springer, New York, chapter 1, 1–28.

[23] Koonin, EV (2012). *The Logic of Chance: The Nature and Origin of Biological Evolution*. FT Press Science, Upper Saddle River, NJ.

[24] Soyer, O and O'Malley, M (2013). Evolutionary systems biology: What it is and why it matters. *BioEssays*, **35**:696–705.

[25] Rowe, W, Platt, M, Wedge, DC, Day, PJ, Kell, DB, and Knowles, J (2010). Analysis of a complete DNA-protein affinity landscape. *Journal of the Royal Society Interface*, **7**:397–408.

[26] Hinkley, T, Martins, J, Chappey, C, Haddad, M, and Stawiski et al., E (2011). A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature Genetics*, **43**:487–489.

[27] Jiménez, JI, Xulvi-Brunet, R, Campbell, GW, Turk-MacLeod, R, and Chen, IA (2013). Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proceedings of the National Academy of Sciences of the United States of America*, **110**:14984–14989.

[28] Buenrostro, JD, Araya, CL, Chircus, LM, Layton, CJ, Chang, HY, Snyder, MP, and Greenleaf, WJ (2014). Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature Genetics*, **32**:562–568.

[29] Podgornaia, AI and Laub, MT (2015). Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, **347**:673–677.

[30] Julien, P, Miñana, B, Baeza-Centurion, P, Valcárcel, J, and Lehner, B (2016). The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nature Communications*, **7**:11558.

[31] Li, C, Qian, W, Maclean, CJ, and Zhang, J (2016). The fitness landscape of a tRNA gene. *Science*, **352**:837–840.

[32] Puchta, O, Cseke, B, Czaja, H, Tollervey, D, Sanguinetti, G, and Kudla, G (2016). Network of epistatic interactions within a yeast snoRNA. *Science*, **352**:840–844.

[33] Qiu, C, Erinne, OC, Dave, JM, Cui, P, Jin, H, Muthukrishnan, N, Tang, LK, Babu, SG, Lam, KC, Vandeventer, PJ, Strohner, R, Van den Brulle, J, Sze, SH, and Kaplan, CD (2016). High-resolution phenotypic landscape of the RNA Polymerase II Trigger Loop. *PLOS Genet*, **12**:e1006321.

[34] Sarkisyan, KS, Bolotin, DA, Veer, MV, Usmanova, DR, and Mishin et al., AS (2016). Local fitness landscape of the green fluorescent protein. *Nature*, **533**:397–401.

[35] Alberch, P (1991). From genes to phenotype: dynamical systems and evolvability. *Genetica*, **84**:5–11.

[36] Wagner, GP and Zhang, J (2011). The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, **12**:204–213.

[37] Lehner, B (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*, **14**:168–178.

[38] Raman, K and Wagner, A (2011). The evolvability of programmable hardware. *Journal of the Royal Society Interface*, **8**:269–281.

[39] Fortuna, M, Zaman, L, Ofria, C, and Wagner, A (2017). The genotype-phenotype map of an evolving digital organism. *PLOS Computational Biology*, **13**:e1005414.

[40] Wright, S (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In DF Jones, editor, *Proceedings of the Sixth International Congress on Genetics*. The Genetics Society of America, volume 1, 356–366.

[41] Maynard Smith, J (1970). Natural selection and the concept of a protein space. *Nature*, **225**:563–564.

[42] Burns, J (1970). The synthetic problem and the genotype-phenotype relation in cellular metabolism. In CH Waddington, editor, *Towards a Theoretical Biology. Volume 3: Drafts. An IUBS Symposium*, Aldine Publishing Company, Chicago, 47–51.

[43] Orr, HA (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, **10**:531–539.

[44] Barrick, JE and Lenski, RE (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, **14**:827–839.

[45] Hansen, TF (2016). On the definition and measurement of fitness in finite populations. *Journal of Theoretical Biology*, **419**:36–43.

[46] Provine, WB (1986). *Sewall Wright and Evolutionary Biology*. The University of Chicago Press, Chicago.

[47] Gavrilets, S (2004). *Fitness landscapes and the origing of species*. Princeton University Press, Princeton.

[48] Kondrashov, FA and Kondrashov, AS (2001). Multidimensional epistasis and the disadvantage of sex. *Proceedings of the National Academy of Sciences of the United States of America*, **98**:12089–12092.

[49] de Visser, JAGM, Park, SC, and Krug, J (2009). Exploring the effect of sex on empirical fitness landscapes. *The American Naturalist*, **174 Suppl 1**:S15–S30.

[50] van Nimwegen, E, Crutchfield, JP, and Huynen, M (1999). Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences of the United States of America*, **96**:9716–9720.

[51] Wagner, A (2005). *Robustness and Evolvability in Living Systems*. Princeton University Press.

[52] Salverda, MLM, Dellus, E, Gorter, Fa, Debets, AJM, van der Oost, J, Hoekstra, RF, Tawfik, DS, and de Visser, JAGM (2011). Initial mutations direct alternative pathways of protein evolution. *PLOS Genetics*, **7**:e1001321.

[53] Lobkovsky, AE, Wolf, YI, and Koonin, EV (2011). Predictability of evolutionary trajectories in fitness landscapes. *PLOS Computational Biology*, **7**:e1002302.

[54] Lässig, M, Mustonen, V, and Walczak, AM (2017). Predicting evolution. *Nature Ecology & Evolution*, **1**:0077.

[55] Szendro, IG, Schenk, MF, Franke, J, Krug, J, and de Visser, JAGM (2013). Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, **2013**:P01005.

[56] de Visser, JAGM and Krug, J (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, **15**:480–490.

[57] Kondrashov, DA and Kondrashov, FA (2015). Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics*, **31**:24–33.

[58] Obolski, U, Ram, Y, and Hadany, L (2017). Key Issues Review: Evolution on rugged adaptive landscapes. *bioRxiv*.

[59] Lipman, DJ and Wilbur, JW (1991). Modelling neutral and selective evolution of protein folding. *Proceedings of the Royal Society B*, **245**:7–11.

[60] Schuster, P, Fontana, W, Stadler, PF, and Hofacker, IL (1994). From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society B*, **255**:279–284.

[61] Ciliberti, S, Martin, OC, and Wagner, A (2007). Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, **104**:13591–13596.

[62] Rodrigues, JFM and Wagner, A (2009). Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLOS Computational Biology*, **5**:e1000613.

[63] Arias, CF, Catalán, P, Manrubia, S, and Cuesta, Ja (2014). toyLIFE: a computational framework to study the multi-level organisation of the genotype-phenotype map. *Scientific Reports*, **4**:7549.

[64] Karr, JR, Sanghvi, JC, MacKlin, DN, Gutschow, MV, Jacobs, JM, Bolival, B, Assad-Garcia, N, Glass, JI, and Covert, MW (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, **150**:389–401.

[65] Salazar-Ciudad, I and Jernvall, J (2010). A computational model of teeth and the developmental origins of morphological variation. *Nature*, **464**:583–586.

[66] Salazar-Ciudad, I and Marín-Riera, M (2013). Adaptive dynamics under development-based genotype-phenotype maps. *Nature*, **497**:361–364.

[67] Draghi, JA, Parsons, TL, Wagner, GP, and Plotkin, JB (2010). Mutational robustness can facilitate adaptation. *Nature*, **463**:353–355.

[68] Manrubia, S and Cuesta, JA (2015). Evolution on neutral networks accelerates the ticking rate of the molecular clock. *Journal of the Royal Society Interface*, **12**:20141010.

[69] Wagner, A (2008). Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B*, **275**:91–100.

[70] Cowperthwaite, MC, Economo, EP, Harcombe, WR, Miller, EL, and Meyers, LA (2008). The ascent of the abundant: How mutational networks constrain evolution. *PLOS Computational Biology*, **4**:e1000110.

[71] McCandlish, DM (2013). On the findability of genotypes. *Evolution*, **67**:2592–2603.

[72] Schaper, S and Louis, AA (2014). The arrival of the frequent: How bias in genotype-phenotype maps can steer populations to local optima. *PLOS ONE*, **9**:e86635.

[73] Venkataram, S, Dunn, B, Li, Y, Agarwala, A, Chang, J, Ebel, ER, Geiler-Samerotte, K, Hérissant, L, Blundell, JR, Levy, SF, Fisher, DS, Sherlock, G, and Petrov, DA (2016). Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast. *Cell*, **166**:1585–1596.

[74] Aguilar-Rodríguez, J, Payne, JL, and Wagner, A (2017). A thousand adaptive landscapes and their navigability. *Nature Ecology & Evolution*, **1**:0045.

[75] Blundell, JR and Levy, SF (2014). Beyond genome sequencing: lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. *Genomics*, **104**:417–430.

[76] Long, A, Liti, G, Luptak, A, and Tenaillon, O (2015). Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nature Reviews Genetics*, **16**:567–582.

[77] Khalid, F, Aguilar-Rodríguez, J, Wagner, A, and Payne, JL (2016). Genonets server-a web server for the construction, analysis and visualization of genotype networks. *Nucleic Acids Research*, **44**:W70–6.

[78] Wagner, A (2014). Mutational robustness accelerates the origin of novel RNA phenotypes through phenotypic plasticity. *Biophysical Journal*, **106**:955–965.

[79] Dawid, A, Kiviet, DJ, Kogenaru, M, de Vos, M, and Tans, SJ (2010). Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape. *Chaos*, **20**:026105.

[80] Phillips, PC (2008). Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9**:855–67.

[81] de Visser, JAGM, Cooper, TF, and Elena, SF (2011). The causes of epistasis. *Proceedings of the Royal Society B*, **278**:3617–3624.

[82] Poelwijk, FJ, Kiviet, DJ, Weinreich, DM, and Tans, SJ (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, **445**:383–386.

[83] Lehner, B (2013). Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, **27**:323–31.

[84] Ancel, LW and Fontana, W (2000). Plasticity, evolvability, and modularity in RNA. *The Journal of Experimental Zoology*, **288**:242–283.

[85] Kern, AD and Kondrashov, FA (2004). Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. *Nature Genetics*, **36**:1207–1212.

[86] Lunzer, M, Miller, SP, Felsheim, R, and Dean, AM (2005). The biochemical architecture of an ancient adaptive landscape. *Science*, **310**:499–501.

[87] Bershtein, S, Segal, M, Bekerman, R, Tokuriki, N, and Tawfik, DS (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, **444**:929–932.

[88] Weinreich, DM, Delaney, NF, Depristo, MA, and Hartl, DL (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, **312**:111–114.

[89] Ortlund, EA, Bridgham, JT, Redinbo, MR, and Thornton, JW (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, **317**:1544–1548.

[90] Starr, TN and Thornton, JW (2016). Epistasis in protein evolution. *Protein Science*, **25**:1204–1218.

[91] Segrè, D, Deluna, A, Church, GM, and Kishony, R (2005). Modular epistasis in yeast metabolism. *Nature Genetics*, **37**:77–83.

[92] He, X, Qian, W, Wang, Z, Li, Y, and Zhang, J (2010). Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nature Genetics*, **42**:272–276.

[93] Snitkin, ES and Segrè, D (2011). Epistatic interaction maps relative to multiple metabolic phenotypes. *PLOS Genetics*, **7**:e1001294.

[94] Macía, J, Solé, RV, and Elena, SF (2012). The causes of epistasis in genetic networks. *Evolution*, **66**:586–596.

[95] Svensson, E and Calsbeek, R, editors (2012). *The Adaptive Landscape in Evolutionary Biology*. Oxford University Press, Oxford.

[96] Kauffman, S and Levin, S (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, **128**:11–45.

[97] Fisher, RA (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

[98] Wright, S (1931). Evolution in Mendelian populations. *Genetics*, **16**:97–159.

[99] Haldane, JBS (1932). *The Causes of Evolution*. Longmans Green, London.

[100] Ridley, M (2003). *Evolution*. Wiley-Blackwell Publishing, New York.

[101] Barton, NH, Briggs, DEG, Eisen, JA, Goldstein, DB, and Patel, N (2007). *Evolution*. Cold Spring Harbor Laboratory Press, New York.

[102] Freeman, S and Herron, J (2007). *Evolutionary Analysis*. Benjamin Cummins, New York, 4th edition.

[103] Futuyma, D (2009). *Evolution*. Sinauer Associates, Sunderland, MA, 2nd edition.

[104] Dietrich, MR and Skipper, Jr., RA (2012). A Shifting Terrain: A Brief History of the Adaptive Landscape. In E Svensson and R Calsbeek, editors, *The Adaptive Landscape in Evolutionary Biology*, Oxford University Press, Oxford, 3–15.

[105] Haldane, JBS (1931). A Mathematical Theory of Natural Selection. Part VIII. Metastable Populations. *Mathematical Proceedings of the Cambridge Philosophical Society*, **27**:137–142.

[106] Fisher, RA (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics*, **11**:53–63.

[107] Huxley, J (2009). *Evolution: The Modern Synthesis*. The MIT Press, Cambridge, MA, the definitive edition.

[108] Simpson, GG (1944). *Tempo and Mode in Evolution*. Columbia University Press, New York.

[109] McGhee, G (2007). *The Geometry of Evolution: Adaptive Landscapes and Theoretical Morphospaces*. Cambridge University Press, Cambridge, UK.

[110] Pigliucci, M (2012). Landscapes, Surfaces, and Morphospaces: What Are They Good For? In E Svensson and R Calsbeek, editors, *The Adaptive Landscape in Evolutionary Biology*, Oxford University Press, Oxford, 26–38.

[111] Raup, DM (1966). Geometric analysis of shell coiling: General problems. *Journal of Paleontology*, **40**:1178–1190.

[112] Raup, DM (1967). Geometric analysis of shell coiling: Coiling in Ammonoids. *Journal of Paleontology*, **41**:43–65.

[113] Chamberlain, JA (1976). Flow patterns and drag coefficients of cephalopod shells. *Palaeontology*, **19**:539–563.

[114] Chamberlain, JA (1981). Hydromechanical design of fossil cephalopods. *Systematics Association Special Volume*, **18**:289–336.

[115] Saunders, WB, Work, DM, and Nikolaeva, SV (2004). Morphology and morphologic diversity of mid-Carboniferous (Namurian) ammonoids in time and space. *Paleobiology*, **10**:195–228.

[116] Judson, H (2013). *The eight day of creation: Makers of the revolution in biology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, commemorative edition.

[117] Cobb, M (2015). *Life's Greatest Secret: The Race to Crack the Genetic Code*. Profile Books, London.

[118] Anfinsen, C (1959). *The Molecular Basis of Evolution*. John wiley and Sons, Inc., New York.

[119] Zuckerkandl, E and Pauling, L (1965). Evolutionary divergence and convergence in proteins. In V Bryson and H Vogel, editors, *Evolving Genes and Proteins*, Academic Press, New York, 97–166.

[120] Jukes, T (1966). *Molecules and Evolution*. Columbia University Press, New York.

[121] Maynard Smith, J (1962). The limitations of molecular evolution. In IJ Good, editor, *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, Basic Books, New York, 252–256.

[122] Greenbury, SF, Schaper, S, Ahnert, SE, and Louis, AA (2016). Genetic correlations greatly increase mutational robustness and can both reduce and enhance evolvability. *PLOS Computat Biol*, **12**:e1004773.

[123] Gillespie, JH (2011). Evolution. *Bioinformatics*, **27**:1017–1018.

[124] Gillespie, JH (1991). *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

[125] Aita, T and Husimi, Y (1996). Fitness spectrum among random mutants on Mt. Fuji-type fitness landscape. *Journal of Theoretical Biology*, **182**:469–485.

[126] Kingman, JFC (1977). Properties of bilinear models for balance between genetic mutation and selection. *Mathematical Proceedings of the Cambridge Philosophical Society*, **81**:443–453.

[127] Kingman, JFC (1978). A simple model for the balance between selection and mutation. *Journal of Applied Probability*, **15**:1–12.

[128] Aita, T, Uchiyama, H, Inaoka, T, Nakajima, M, Kokubo, T, and Husimi, Y (2000). Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin. *Biopolymers*, **54**:64–79.

[129] Neidhart, J, Szendro, IG, and Krug, J (2014). Adaptation in tunably rugged fitness landscapes: The rough Mount Fuji model. *Genetics*, **198**:699–713.

[130] Kauffman, SA and Weinberger, ED (1989). The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, **141**(2):211–245.

[131] Kauffman, S (1993). *The Origins of Order*. Oxford University Press, Oxford.

[132] Kauffman, S (1995). *At Home in the Universe*. Oxford University Press, Oxford.

[133] Fontana, W, Stadler, PF, Bornberg-Bauer, EG, Griesmacher, T, Hofacker, IL, Tacker, M, Tarazona, P, Weinberger, ED, and Schuster, P (1993). RNA folding and combinatory landscapes. *Physical Review E*, **47**:2083–2099.

[134] Jones, T (1995). *Evolutionary Algorithms, Fitness Landscapes and Search*. Ph.D. thesis, The University of New Mexico.

[135] Hartl, DL (2014). What can we learn from fitness landscapes? *Current Opinion in Microbiology*, **21**:51–57.

[136] Dobzhansky, T (1970). *Genetics of the Evolutionary Process*. Columbia University Press, New York.

[137] Wagner, A (2014). *Arrival of the Fittest: Solving Evolution's Greatest Puzzle*. Current, New York.

[138] Hayashi, Y, Aita, T, Toyota, H, Husimi, Y, Urabe, I, and Yomo, T (2006). Experimental rugged fitness landscape in protein sequence space. *PLOS ONE*, **1**:e96.

[139] Romero, PA and Arnold, FH (2009). Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, **10**:866–876.

[140] Carneiro, M and Hartl, DL (2010). Colloquium papers: Adaptive landscapes and protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **107 Suppl**:1747–1751.

[141] Romero, PA, Krause, A, and Arnold, FH (2013). Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences of the United States of America*, **110**:E193–E201.

[142] Firnberg, E, Labonte, JW, Gray, JJ, and Ostermeier, M (2014). A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular Biology and Evolution*, **31**:1581–1592.

[143] Gupta, A and Adami, C (2016). Strong selection significantly increases epistatic interactions in the long-term evolution of a protein. *PLOS Genetics*, **12**:e1005960.

[144] Wu, NC, Dai, L, Olson, CA, Lloyd-Smith, JO, and Sun, R (2016). Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, **5**:1–21.

[145] Lee, YH, DSouza, LM, and Fox, GE (1997). Equally parsimonious pathways through an RNA sequence space are not equally likely. *Journal of Molecular Evolution*, **45**:278–284.

[146] Pitt, JN and Ferré-D'Amaré, AR (2010). Rapid construction of empirical RNA fitness landscapes. *Science*, **330**(6002):376–379.

[147] Athavale, SS, Spicer, B, and Chen, IA (2014). Experimental fitness landscapes to understand the molecular evolution of RNA-based life. *Current Opinion in Chemical Biology*, **22**:35–39.

[148] Weinreich, DM, Watson, RA, and Chao, L (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, **59**:1165–1174.

[149] Poelwijk, FJ, Tănase-Nicola, S, Kiviet, DJ, and Tans, SJ (2011). Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of Theoretical Biology*, **272**:141–144.

[150] Grüner, W, Giegerich, R, Strothmann, D, Reidys, C, Weber, J, Hofacker, IL, Stadler, PF, and Schuster, P (1996). Analysis of RNA sequence structure maps by exhaustive enumeration I. Neutral networks. *Monatshefte für Chemie/Chemical Monthly*, **127**:355–374.

[151] Reidys, C, Stadler, PF, and Schuster, P (1997). Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bulletin of Mathematical Biology*, **59**:339–397.

[152] Stadler, BMR, Stadler, PF, Wagner, GP, and Fontana, W (2001). The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology*, **213**:241–274.

[153] Aguirre, J, Buldú, J, Stich, M, and Manrubia, S (2011). Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLOS ONE*, **6**:e26324.

[154] Ferrada, E and Wagner, A (2012). A comparison of genotype-phenotype maps for RNA and proteins. *Biophysical Journal*, **102**:1916–1925.

[155] Cotterell, J and Sharpe, J (2010). An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Molecular Systems Biology*, **6**:425.

[156] Payne, JL and Wagner, A (2014). The robustness and evolvability of transcription factor binding sites. *Science*, **466**:714–719.

[157] Ptashne, M and Gann, A (2002). *Genes & Signals*. Cold Spring Harbor Laboratory Press, New York.

[158] Stormo, GD (2013). *Introduction to Protein-DNA Interactions: Structure, Thermodynamics, and Bioinformatics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

[159] Gertz, J and Cohen, BA (2009). Environment-specific combinatorial cis-regulation in synthetic promoters. *Molecular Systems Biology*, **5**:244.

[160] Shultzaberger, RK, Malashock, DS, Kirsch, JF, and Eisen, MB (2010). The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. *PLOS Genetics*, **6**:e1001042.

[161] Kheradpour, P, Ernst, J, Melnikov, A, Rogov, P, Wang, L, Zhang, X, Alston, J, Mikkelsen, TS, and Kellis, M (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, **23**:800–811.

[162] Weingarten-Gabbay, S and Segal, E (2014). The grammar of transcriptional regulation. *Human Genetics*, **133**:701–711.

[163] Stormo, GD and Zhao, Y (2010). Determining the specificity of protein-DNA interactions. *Nature Reviews Genetics*, **11**:751–760.

[164] Davidson, EH (2006). *The regulatory genome: gene regulatory networks in development and evolution.* Academic Press, San Diego.

[165] Levine, M and Tjian, R (2003). Transcription regulation and animal diversity. *Nature*, **424**:147–151.

[166] Prud'homme, B, Gompel, N, and Carroll, SB (2007). Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **104**:8605–8612.

[167] Carroll, SB (2008). Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, **134**:25–36.

[168] Schaub, MA, Boyle, AP, Kundaje, A, Batzoglou, S, and Snyder, M (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*, **22**:1748–1759.

[169] Ward, LD and Kellis, M (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, **30**:1095–1106.

[170] Gerstein, MB, Kundaje, A, Hariharan, M, Landt, SG, Yan, KK, Cheng, C, Mu, XJ, Khurana, E, Rozowsky, J, Alexander, R, Min, R, Alves, P, Abyzov, A, Addleman, N, Bhardwaj, N, Boyle, AP, Cayting, P, Charos, A, Chen, DZ, Cheng, Y, Clarke, D, Eastman, C, Euskirchen, G, Frietze, S, Fu, Y, Gertz, J, Grubert, F, Harmanci, A, Jain, P, Kasowski, M, Lacroute, P, Leng, J, Lian, J, Monahan, H, O'Geen, H, Ouyang, Z, Partridge, EC, Patacsil, D, Pauli, F, Raha, D, Ramirez, L, Reddy, TE, Reed, B, Shi, M, Slifer, T, Wang, J, Wu, L, Yang, X, Yip, KY, Zilberman-Schapira, G, Batzoglou, S, Sidow, A, Farnham, PJ, Myers, RM, Weissman, SM, and Snyder, M (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**:91–100.

[171] Neph, S, Stergachis, AB, Reynolds, A, Sandstrom, R, Borenstein, E, and Stamatoyannopoulos, JA (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, **150**:1274–1286.

[172] Bornberg-Bauer, E and Albà, MM (2013). Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology*, **23**:459–466.

[173] Toll-Riera, M and Albà, MM (2013). Emergence of novel domains in proteins. *BMC Evolutionary Biology*, **13**:47.

[174] Weirauch, MT and Hughes, TR (2011). A Catalogue of Eukaryotic Transcription Factor Types, Their Evolutionary Origin, and Species Distribution. In TR Hughes, editor, *A Handbook of Transcription Factors*, Springer, chapter 3, 25–73.

[175] Rohs, R, West, SM, Sosinsky, A, Liu, P, Mann, RS, and Honig, B (2009). The role of DNA shape in protein-DNA recognition. *Nature*, **461**:1248–1253.

[176] Weirauch, MT, Cote, A, Norel, R, Annala, M, Zhao, Y, Riley, TR, Saez-Rodriguez, J, Cokelaer, T, Vedenko, A, Talukder, S, Bussemaker, HJ, Morris, QD, Bulyk, ML, Stolovitzky, G, and Hughes, TR (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, **31**:126–134.

[177] Stormo, GD, Schneider, TD, and Gold, L (1986). Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Research*, **14**:6661–6679.

[178] Benos, PV, Bulyk, ML, and Stormo, GD (2002). Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Research*, **30**:4442–4451.

[179] Lee, MLT, Bulyk, ML, Whitmore, Ga, and Church, GM (2002). A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. *Biometrics*, **58**:981–988.

[180] Barash, Y, Elidan, G, Friedman, N, and Kaplan, T (2003). Modeling dependencies in protein-DNA binding sites. *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, **21**:28–37.

[181] King, OD and Roth, FP (2003). A non-parametric model for transcription factor binding sites. *Nucleic Acids Research*, **31**:e116.

[182] Zhou, Q and Liu, JS (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**:909–916.

[183] Stormo, GD and Zhao, Y (2007). Putting numbers on the network connections. *BioEssays*, **29**:717–721.

[184] Berger, MF, Philippakis, AA, Qureshi, AM, He, FS, Estep, PW, and Bulyk,

ML (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, **24**:1429–1435.

[185] Warren, CL, Kratochvil, NCS, Hauschild, KE, Foister, S, Brezinski, ML, Dervan, PB, Phillips, GN, and Ansari, AZ (2006). Defining the sequence-recognition profile of DNA-binding molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **103**:867–872.

[186] Maerkl, SJ and Quake, SR (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, **315**:233–237.

[187] Fordyce, PM, Gerber, D, Tran, D, Zheng, J, Li, H, DeRisi, JL, and Quake, SR (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature Biotechnology*, **28**:970–975.

[188] Johnson, DS, Mortazavi, A, Myers, RM, and Wold, B (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**:1497–1502.

[189] Nutiu, R, Friedman, RC, Luo, S, Khrebtukova, I, Silva, D, Li, R, Zhang, L, Schroth, GP, and Burge, CB (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotechnology*, **29**:659–664.

[190] Sharon, E, Kalma, Y, Sharp, A, Raveh-Sadka, T, Levo, M, Zeevi, D, Keren, L, Yakhini, Z, Weinberger, A, and Segal, E (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, **30**:521–530.

[191] Patwardhan, RP, Hiatt, JB, Witten, DM, Kim, MJ, Smith, RP, May, D, Lee, C, Andrie, JM, Lee, SI, Cooper, GM, Ahituv, N, Pennacchio, LA, and Shendure, J (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, **30**:265–270.

[192] Bulyk, ML, Huang, X, Choo, Y, and Church, GM (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, **98**:7158–7163.

[193] Mukherjee, S, Berger, MF, Jona, G, Wang, XS, Muzzey, D, Snyder, M, Young, RA, and Bulyk, ML (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics*, **36**:1331–1339.

[194] Zhu, C, Byers, KJRP, McCord, RP, Shi, Z, Berger, MF, Newburger, DE, Saulrieta, K, Smith, Z, Shah, MV, Radhakrishnan, M, Philippakis, Aa, Hu, Y, De Masi, F, Pacek, M, Rolfs, A, Murthy, T, Labaer, J, and Bulyk, ML (2009). High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research*, **19**:556–566.

[195] Badis, G, Berger, MF, Philippakis, AA, Talukder, S, Gehrke, AR, Jaeger, SA, Chan, ET, Metzler, G, Vedenko, A, Chen, X, Kuznetsov, H, Wang, C, Coburn, D, Newburger, DE, Morris, Q, Hughes, TR, and Bulyk, ML (2009). Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**:1720–1723.

[196] Weirauch, MT, Yang, A, Albu, M, Cote, AG, Montenegro-Montero, A, Drewe, P, Najafabadi, HS, Lambert, SA, Mann, I, Cook, K, Zheng, H, Goity, A, van Bakel, H, Lozano, JC, Galli, M, Lewsey, MG, Huang, E, Mukherjee, T, Chen, X, Reece-Hoyes, JS, Govindarajan, S, Shaulsky, G, Walhout, AJM, Bouget, FY, Ratsch, G, Larrondo, LF, Ecker, JR, and Hughes, TR (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**:1431–1443.

[197] Berger, MF and Bulyk, ML (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols*, **4**:393–411.

[198] Puckett, JW, Muzikar, KA, Tietjen, J, Warren, CL, Ansari, AZ, and Dervan, PB

(2007). Quantitative microarray profiling of DNA-binding molecules. *Journal of the American Chemical Society*, **129**:12310–12319.

[199] Carlson, CD, Warren, CL, Hauschild, KE, Ozers, MS, Qadir, N, Bhimsaria, D, Lee, Y, Cerrina, F, and Ansari, AZ (2010). Specificity landscapes of DNA binding molecules elucidate biological function. *Proceedings of the National Academy of Sciences of the United States of America*, **107**:4544–4549.

[200] O'Flanagan, RA, Paillard, G, Lavery, R, and Sengupta, AM (2005). Non-additivity in protein-DNA binding. *Bioinformatics*, **21**:2254–2263.

[201] Zhao, Y and Stormo, GD (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, **29**:480–483.

[202] Morris, Q, Bulyk, ML, and Hughes, TR (2011). Jury remains out on simple models of transcription factor specificity. *Nature Biotechnology*, **29**:483–484.

[203] Stormo, GD (2000). DNA binding sites: representation and discovery. *Bioinformatics*, **16**:16–23.

[204] D'haeseleer, P (2006). What are DNA sequence motifs? *Nature Biotechnology*, **24**:423–425.

[205] Schell, R, Mullis, M, and Ehrenreich, IM (2016). Modifiers of the Genotype–Phenotype Map: Hsp90 and Beyond. *PLOS Biology*, **14**:e2001015.

[206] Fisher, RA (1928). The possible modification of the response of the wildtype to recurrent mutation. *The American Naturalist*, **62**:115–126.

[207] Fisher, RA (1931). The evolution of dominance. *Biological Reviews*, **62**:345–368.

[208] Wright, S (1934). Molecular and evolutionary theories of dominance. *The American Naturalist*, **63**:24–53.

[209] Crosby, J (1963). The evolution and nature of dominance. *Journal of Theoretical Biology*, **5**:35–51.

[210] Charlesworth, B (1979). Evidence against Fihser's theory of dominance. *Nature*, **278**:848–849.

[211] Ellis, J (1987). Proteins as molecular chaperones. *Nature*, **328**:378–379.

[212] Rutherford, SL (2003). Between genotype and phenotype: protein chaperones and evolvability. *Nature Reviews Genetics*, **4**:263–274.

[213] Hartl, FU and Hayer-Hartl, M (2009). Converging concepts of protein folding in vitro and in vivo. *Nature Structural & Molecular Biology*, **16**:574–81.

[214] Hartl, FU, Bracher, A, and Hayer-Hartl, M (2011). Molecular chaperones in protein folding and proteostasis. *Nature*, **475**:324–332.

[215] Lin, Z, Madan, D, and Rye, HS (2008). GroEL stimulates protein folding through forced unfolding. *Nature Structural & Molecular Biology*, **15**:303–311.

[216] Bogumil, D and Dagan, T (2012). Cumulative impact of chaperone-mediated folding on genome evolution. *Biochemistry*, **51**:9941–9953.

[217] Colon, W and Kelly, JW (1992). Partial denaturation of transthyretin is sufficient for amyloid fibril formation *in vitro*. *Biochemistry*, **31**:8654–8660.

[218] Bucciantini, M, Giannoni, E, Chiti, F, Baroni, F, Formigli, L, Zurdo, J, Taddei, N, Ramponi, G, Dobson, CM, and Stefani, M (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, **416**:507–511.

[219] Geiler-Samerotte, Ka, Dion, MF, Budnik, Ba, Wang, SM, Hartl, DL, and Drummond, DA (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **108**:680–685.

[220] Hardy, J and Allsop, D (1991). Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends in Pharmacological Sciences*, **12**:383–388.

[221] Serrano-Pozo, A, Frosch, MP, Masliah, E, and Hyman, BT (2011). Neuropathological alterations in Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*, **1**:1–23.

[222] Pál, C, Papp, B, and Hurst, LD (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, **158**:927–931.

[223] Krylov, DM, Wolf, YI, Rogozin, IB, and Koonin, EV (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, **13**:2229–2235.

[224] Drummond, DA, Bloom, JD, Adami, C, Wilke, CO, and Arnold, FH (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, **102**:14338–14343.

[225] Drummond, DA and Wilke, CO (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**:341–352.

[226] Pál, C, Papp, B, and Lercher, MJ (2006). An integrated view of protein evolution. *Nature Reviews Genetics*, **7**:337–348.

[227] Rocha, EPC and Danchin, A (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular Biology and Evolution*, **21**:108–116.

[228] Ribas de Pouplana, L, Santos, MaS, Zhu, JH, Farabaugh, PJ, and Javid, B (2014). Protein mistranslation: Friend or foe? *Trends in Biochemical Sciences*, 355–362.

[229] Bratulic, S, Gerber, F, and Wagner, A (2015). Mistranslation drives the evolution of robustness in TEM-1 β-lactamase. *Proceedings of the National Academy of Sciences of the United States of America*, **112**:12758–12763.

[230] Bratulic, S, Toll-Riera, M, and Wagner, A (2017). Mistranslation can enhance fitness through purging of deleterious mutations. *Nature Communications*, **8**:1–22.

[231] Rutherford, SL and Lindquist, S (1998). Hsp90 as a capacitor for morphological evolution. *Nature*, **396**:336–342.

[232] Tomala, K and Korona, R (2008). Molecular chaperones and selection against mutations. *Biology Direct*, **3**:5.

[233] Taipale, M, Jarosz, DF, and Lindquist, S (2010). HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nature Reviews Molecular Cell Biology*, **11**:515–528.

[234] Jarosz, DF, Taipale, M, and Lindquist, S (2010). Protein homeostasis and the phenotypic manifestation of genetic diversity: principles and mechanisms. *Annual Review of Genetics*, **44**:189–216.

[235] Queitsch, C, Sangster, TA, and Lindquist, S (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, **417**:618–624.

[236] Yeyati, PL, Bancewicz, RM, Maule, J, and Van Heyningen, V (2007). Hsp90 selectively modulates phenotype in vertebrate development. *PLOS Genetics*, **3**:e42.

[237] Sangster, TA, Salathia, N, Lee, HN, Watanabe, E, Schellenberg, K, Morneau, K, Wang, H, Undurraga, S, Queitsch, C, and Lindquist, S (2008). HSP90-buffered genetic variation is common in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, **105**:2969–2974.

[238] Rohner, N, Jarosz, DF, Kowalko, JE, Yoshizawa, M, Jeffery, WR, Borowsky, RL, Lindquist, S, and Tabin, CJ (2013). Cryptic variation in morphological evolution: HSP90 as a capacitor for loss of eyes in cavefish. *Science*, **342**:1372–1375.

[239] Geiler-Samerotte, KA, Zhu, YO, Goulet, BE, Hall, DW, and Siegal, ML (2016).

Selection Transforms the Landscape of Genetic Variation Interacting with Hsp90. *PLOS Biology*, **14**:e2000465.

[240] Calloni, G, Chen, T, Schermann, SM, Chang, HC, Genevaux, P, Agostini, F, Tartaglia, GG, Hayer-Hartl, M, and Hartl, FU (2012). DnaK functions as a central hub in the *E. coli* chaperone network. *Cell Reports*, **1**:251–264.

[241] Bukau, B and Walker, GC (1989). Cellular defects caused by deletion of the *Escherichia coli dnaK* gene indicate roles for heat shock protein in normal metabolism. *Journal of Bacteriology*, **171**:2337–2346.

[242] Hansen, S, Lewis, K, and Vulić, M (2008). Role of global regulators and nucleotide metabolism in antibiotic tolerance in *Escherichia coli*. *Antimicrobial Agents and Chemotherapy*, **52**:2718–2726.

[243] Liu, A, Tran, L, Becket, E, Lee, K, Chinn, L, Park, E, Tran, K, and Miller, JH (2010). Antibiotic sensitivity profiles determined with an *Escherichia coli* gene knockout collection: Generating an antibiotic bar code. *Antimicrobial Agents and Chemotherapy*, **54**:1393–1403.

[244] Straus, D, Walter, W, and Gross, Ca (1990). DnaK, DnaJ, and GrpE heat shock proteins negatively regulate heat shock gene expression by controlling the synthesis and stability of sigma 32. *Genes & Development*, **4**:2202–2209.

[245] Hoffmann, HJ, Lyman, SK, Lu, C, Petit, MA, and Echols, H (1992). Activity of the Hsp70 chaperone complex–DnaK, DnaJ, and GrpE–in initiating phage lambda DNA replication by sequestering and releasing lambda P protein. *Proceedings of the National Academy of Sciences of the United States of America*, **89**:12108–12111.

[246] Rüdiger, S, Buchberger, A, and Bukau, B (1997). Interaction of Hsp70 chaperones with substrates. *Nature Structural Biology*, **4**:342–349.

[247] Clare, DK, Vasishtan, D, Stagg, S, Quispe, J, Farr, GW, Topf, M, Horwich, AL,

and Saibil, HR (2012). ATP-triggered conformational changes delineate substrate-binding and -folding mechanics of the GroEL chaperonin. *Cell*, **149**:113–123.

[248] Sigler, PB, Xu, Z, Rye, HS, Burston, SG, Fenton, WA, and Horwich, AL (1998). Structure and function in GroEL-mediated protein folding. *Annual Review of Biochemistry*, **67**:581–608.

[249] Kerner, MJ, Naylor, DJ, Ishihama, Y, Maier, T, Chang, HC, Stines, AP, Georgopoulos, C, Frishman, D, Hayer-Hartl, M, Mann, M, and Hartl, FU (2005). Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. *Cell*, **122**:209–220.

[250] Fujiwara, K, Ishihama, Y, Nakahigashi, K, Soga, T, and Taguchi, H (2010). A systematic survey of in vivo obligate chaperonin-dependent substrates. *The EMBO Journal*, **29**:1552–1564.

[251] Wang, JD, Herman, C, Tipton, Ka, Gross, Ca, and Weissman, JS (2002). Directed evolution of substrate-optimized GroEL/S chaperonins. *Cell*, **111**:1027–39.

[252] Moya, A, Peretó, J, Gil, R, and Latorre, A (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nature Reviews Genetics*, **9**:218–229.

[253] Toft, C and Andersson, SGE (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature Reviews Genetics*, **11**:465–475.

[254] McCutcheon, JP and Moran, NA (2012). Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, **10**:13–26.

[255] Latorre, A and Manzano-Marín, A (2017). Dissecting genome reduction and trait loss in insect endosymbionts. *Annals of the New York Academy of Sciences*, **1389**:52–75.

[256] Muller, H (1964). The relation of recombination to mutational advance. *Mutation Research*, **1**(1):2–9.

[257] Moran, NA (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **93**:2873–2878.

[258] Fares, MA, Ruiz-González, MX, Moya, A, Elena, SF, and Barrio, E (2002). Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature*, **417**:398.

[259] Baumann, P, Baumann, L, and Clark, MA (1996). Levels of *Buchnera aphidicola* chaperonin Groel during growth of the aphid *Schizaphis graminum*. *Current Microbiology*, **32**:279–285.

[260] Wilcox, JL, Dunbar, HE, Wolfinger, RD, and Moran, NA (2003). Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Molecular Microbiology*, **48**:1491–1500.

[261] McCutcheon, JP, McDonald, BR, and Moran, NA (2009). Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLOS Genetics*, **5**.

[262] Stoll, S, Feldhaar, H, and Gross, R (2009). Transcriptional profiling of the endosymbiont *Blochmannia floridanus* during different developmental stages of its holometabolous ant host. *Environmental Microbiology*, **11**:877–888.

[263] Bennett, GM and Moran, NA (2013). Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biology and Evolution*, **5**:1675–1688.

[264] Fan, Y, Thompson, JW, Dubois, LG, Moseley, MA, and Wernegreen, JJ (2013). Proteomic analysis of an unculturable bacterial endosymbiont (*Blochmannia*) reveals high abundance of chaperonins and biosynthetic enzymes. *Journal of Proteome Research*, **12**:704–718.

[265] Oakeson, KF, Gil, R, Clayton, AL, Dunn, DM, Von Niederhausern, AC, Hamil, C, Aoyagi, A, Duval, B, Baca, A, Silva, FJ, Vallier, A, Jackson, DG, Latorre, A, Weiss, RB, Heddi, A, Moya, A, and Dale, C (2014). Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biology and Evolution*, **6**:76–93.

[266] Sabater-Muñoz, B, Prats-Escriche, M, Montagud-Martínez, R, López-Cerdán, A, Toft, C, Aguilar-Rodríguez, J, Wagner, A, and Fares, MA (2015). Fitness Trade-Offs Determine the Role of the Molecular Chaperonin GroEL in Buffering Mutations. *Molecular Biology and Evolution*, **32**:2681–2693.

[267] Maisnier-Patin, S, Roth, JR, Fredriksson, A, Nyström, T, Berg, OG, and Andersson, DI (2005). Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nature Genetics*, **37**:1376–1379.

[268] Tokuriki, N and Tawfik, DS (2009). Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology*, **19**:596–604.

[269] Gibson, G and Dworkin, I (2004). Uncovering cryptic genetic variation. *Nature Reviews Genetics*, **5**:681–690.

[270] Amitai, G, Gupta, RD, and Tawfik, DS (2007). Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP journal*, **1**:67–78.

[271] Le Rouzic, A and Carlborg, Ö (2008). Evolutionary potential of hidden genetic variation. *Trends in Ecology and Evolution*, **23**:33–37.

[272] Hayden, EJ, Ferrada, E, and Wagner, A (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature*, **474**:92–95.

[273] Wilke, CO, Bloom, JD, Drummond, DA, and Raval, A (2005). Predicting the tolerance of proteins to random amino acid substitution. *Biophysical Journal*, **89**:3714–3720.

[274] DePristo, Ma, Weinreich, DM, and Hartl, DL (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics*, **6**:678–687.

[275] Yue, P, Li, Z, and Moult, J (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, **353**:459–473.

[276] Tokuriki, N, Stricher, F, Serrano, L, and Tawfik, DS (2008). How protein stability and new functions trade off. *PLOS Computational Biology*, e1000002.

[277] Lindner, AB, Madden, R, Demarez, A, Stewart, EJ, and Taddei, F (2008). Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proceedings of the National Academy of Sciences*, **105**:3076–3081.

[278] Tokuriki, N and Tawfik, DS (2009). Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature*, **459**:668–673.

[279] Williams, TA and Fares, MA (2010). The effect of chaperonin buffering on protein evolution. *Genome Biology and Evolution*, 609–619.

[280] Bogumil, D and Dagan, T (2010). Chaperonin-dependent accelerated substitution rates in prokaryotes. *Genome Biology and Evolution*, 602–608.

[281] Aguilar-Rodríguez, J, Sabater-Muñoz, B, Montagud-Martínez, R, Berlanga, V, Alvarez-Ponce, D, Wagner, A, and Fares, MA (2016). The molecular chaperone DnaK is a source of mutational robustness. *Genome Biology and Evolution*, **8**:2979–2991.

[282] Kadibalban, AS, Bogumil, D, Landan, G, and Dagan, T (2016). DnaK-dependent accelerated evolutionary rate in prokaryotes. *Genome Biology and Evolution*, **8**:1590–1599.

[283] Berg, Jeremy M and Tymoczko, John L and Stryer, L (2010). *Biochemestry, 7th edition.* Freeman and Company, New York.

[284] Wagner, A (2012). Metabolic networks and their evolution. In OS Soyer, editor, *Evolutionary Systems Biology*, Springer, New York, chapter 2, 29–52.

[285] Feist, AM, Henry, CS, Reed, JL, Krummenacker, M, Joyce, AR, Karp, PD, Broadbelt, LJ, Hatzimanikatis, V, and Palsson, BØ (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, **3**:121.

[286] Barve, A and Wagner, A (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature*, **500**:203–206.

[287] Feist, AM, Herrgård, MJ, Thiele, I, Reed, JL, and Palsson, BØ (2009). Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, **7**:129–143.

[288] Butcher, EC, Berg, EL, and Kunkel, EJ (2004). Systems biology in drug discovery. *Nature Biotechnology*, **22**:1253–1259.

[289] Lee, JW, Kim, TY, Jang, YS, Choi, S, and Lee, SY (2011). Systems metabolic engineering for chemicals and materials. *Trends in Biotechnology*, **29**:370–378.

[290] Barve, A, Rodrigues, JFM, and Wagner, A (2012). Superessential reactions in metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, **109**:E1121–E1130.

[291] Csete, M and Doyle, J (2004). Bow ties, metabolism and disease. *Trends in Biotechnology*, **22**:446–450.

[292] Friedlander, T, Mayo, AE, Tlusty, T, and Alon, U (2015). Evolution of bow-tie architectures in biology. *PLOS Computational Biology*, **11**:1–19.

[293] Blank, LM, Kuepfer, L, and Sauer, U (2005). Large-scale ^{13}C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biology*, **6**:R49.

[294] Blank, LM, Lehmbeck, F, and Sauer, U (2005). Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. *FEMS Yeast Research*, **5**:545–558.

[295] Fischer, E and Sauer, U (2005). Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nature Genetics*, **37**:636–640.

[296] Orth, JD, Thiele, I, and Palsson, BØ (2010). What is flux balance analysis? *Nature Biotechnology*, **28**:245–248.

[297] Kauffman, KJ, Prakash, P, and Edwards, JS (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, **14**:491–496.

[298] Price, ND, Reed, JL, and Palsson, BØ (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, **2**:886–897.

[299] Palsson, BØ (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, Cambridge, UK.

[300] Bordbar, A, Monk, JM, King, ZA, and Palsson, BØ (2014). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, **15**:107–120.

[301] Varma, A and Palsson, BØ (1993). Metabolic capabilities of *Escherichia coli*: I. synthesis of biosynthetic precursors and cofactors. *Journal of Theoretical Biology*, **165**:477–502.

[302] Schellenberger, J and Palsson, BØ (2009). Use of randomized sampling for analysis of metabolic networks. *Journal of Biological Chemistry*, **284**:5457–5461.

[303] Kaufman, DE and Smith, RL (1998). Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, **46**:84–95.

[304] Bordbar, A, Lewis, NE, Schellenberger, J, Palsson, BØ, and Jamshidi, N (2010). Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Molecular Systems Biology*, **6**:422.

[305] Lewis, NE, Schramm, G, Bordbar, A, Schellenberger, J, Andersen, MP, Cheng, JK, Patel, N, Yee, A, Lewis, Ra, Eils, R, König, R, and Palsson, BØ (2010). Large-scale *in silico* modeling of metabolic interactions between cell types in the human brain. *Nature Biotechnology*, **28**:1279–1285.

[306] Schellenberger, J, Lewis, NE, and Palsson, BØ (2011). Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical Journal*, **100**:544–553.

[307] Nam, H, Lewis, NE, Lerman, JA, Lee, DH, Chang, RL, Kim, D, and Palsson, BØ (2012). Network context and selection in the evolution to enzyme specificity. *Science*, **337**:1101–1104.

[308] Vitkup, D, Kharchenko, P, and Wagner, A (2006). Influence of metabolic network structure and function on enzyme evolution. *Genome Biology*, **7**:R39.

[309] Applebee, MK, Herrgård, MJ, and Palsson, BØ (2008). Impact of individual mutations on increased fitness in adaptively evolved strains of *Escherichia coli*. *Journal of Bacteriology*, **190**:5087–5094.

[310] Blaby, IK, Lyons, BJ, Wroclawska-Hughes, E, Phillips, GCF, Pyle, TP, Chamberlin, SG, Benner, SA, Lyons, TJ, de Crécy-Lagard, V, and de Crécy, E (2012). Experimental evolution of a facultative thermophile from a mesophilic ancestor. *Applied and Environmental Microbiology*, **78**:144–155.

[311] Lee, DH and Palsson, BØ (2010). Adaptive evolution of *Escherichia coli* K-12 MG1655 during growth on a nonnative carbon source, L-l,2-propanediol. *Applied and Environmental Microbiology*, **76**:4158–4168.

[312] Ochman, H, Lawrence, JG, and Groisman, EA (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**:299–304.

[313] Wagner, A (2009). Evolutionary constraints permeate large metabolic networks. *BMC Evolutionary Biology*, **9**:231.

[314] Lerat, E, Daubin, V, Ochman, H, and Moran, NA (2005). Evolutionary origins of genomic repertoires in bacteria. *PLOS Biology*, **3**:e130.

[315] Wray, GA (2007). The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics*, **8**:206–216.

[316] Gertz, J, Siggia, ED, and Cohen, BA (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, **457**:215–218.

[317] Gerland, U and Hwa, T (2002). On the selection and evolution of regulatory DNA motifs. *Journal of Molecular Evolution*, **55**:386–400.

[318] Berg, J, Willmann, S, and Lässig, M (2004). Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology*, **4**:42.

[319] Mustonen, V, Kinney, J, Callan, CG, and Lässig, M (2008). Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences of the United States of America*, **105**:12376–12381.

[320] Haldane, A, Manhart, M, and Morozov, AV (2014). Biophysical fitness landscapes for transcription factor binding sites. *PLOS Computational Biology*, **10**:e1003683.

[321] Weghorn, D and Lässig, M (2013). Fitness landscape for nucleosome positioning. *Proceedings of the National Academy of Sciences of the United States of America*, **110**:10988–10993.

[322] Newburger, DE and Bulyk, ML (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, **37**:D77–82.

[323] Nakagawa, S, Gisselbrecht, SS, Rogers, JM, Hartl, DL, and Bulyk, ML (2013). DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, **110**:12349–12354.

[324] Jolma, A, Yan, J, Whitington, T, Toivonen, J, Nitta, KR, Rastas, P, Morgunova, E, Enge, M, Taipale, M, Wei, G, Palin, K, Vaquerizas, JM, Vincentelli, R, Luscombe, NM, Hughes, TR, Lemaire, P, Ukkonen, E, Kivioja, T, and Taipale, J (2013). DNA-binding specificities of human transcription factors. *Cell*, **152**:327–39.

[325] Yue, F, Cheng, Y, Breschi, A, Vierstra, J, Wu, W, Ryba, T, Sandstrom, R, Ma, Z, Davis, C, Pope, BD, Shen, Y, Pervouchine, DD, Djebali, S, Thurman, RE, Kaul, R, Rynes, E, Kirilusha, A, Marinov, GK, Williams, BA, Trout, D, Amrhein, H, Fisher-Aylor, K, Antoshechkin, I, DeSalvo, G, See, LH, Fastuca, M, Drenkow, J, Zaleski, C, Dobin, A, Prieto, P, Lagarde, J, Bussotti, G, Tanzer, A, Denas, O, Li, K, Bender, MA, Zhang, M, Byron, R, Groudine, MT, McCleary, D, Pham, L, Ye, Z, Kuan, S, Edsall, L, Wu, YC, Rasmussen, MD, Bansal, MS, Kellis, M, Keller, CA, Morrissey, CS, Mishra, T, Jain, D, Dogan, N, Harris, RS, Cayting, P, Kawli, T, Boyle, AP, Euskirchen, G, Kundaje, A, Lin, S, Lin, Y, Jansen, C, Malladi, VS, Cline, MS, Erickson, DT, Kirkup, VM, Learned, K, Sloan, CA, Rosenbloom, KR, Lacerda de Sousa, B, Beal, K, Pignatelli, M, Flicek, P, Lian, J, Kahveci, T, Lee, D, James Kent, W, Ramalho Santos, M, Herrero, J, Notredame, C, Johnson, A, Vong, S, Lee, K, Bates, D, Neri, F, Diegel, M, Canfield, T, Sabo, PJ, Wilken, MS, Reh, TA, Giste, E, Shafer, A, Kutyavin, T, Haugen, E, Dunn, D, Reynolds, AP, Neph, S, Humbert, R, Scott Hansen, R, De Bruijn, M, Selleri, L, Rudensky, A, Josefowicz, S, Samstein, R, Eichler, EE, Orkin, SH, Levasseur, D, Papayannopoulou,

T, Chang, KH, Skoultchi, A, Gosh, S, Disteche, C, Treuting, P, Wang, Y, Weiss, MJ, Blobel, GA, Cao, X, Zhong, S, Wang, T, Good, PJ, Lowdon, RF, Adams, LB, Zhou, XQ, Pazin, MJ, Feingold, EA, Wold, B, Taylor, J, Mortazavi, A, Weissman, SM, Stamatoyannopoulos, JA, Snyder, MP, Guigo, R, Gingeras, TR, Gilbert, DM, Hardison, RC, Beer, MA, and Ren, B (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**:355–364.

[326] Stergachis, AB, Neph, S, Sandstrom, R, Haugen, E, Reynolds, AP, Zhang, M, Byron, R, Canfield, T, Stelhing-Sun, S, Lee, K, Thurman, RE, Vong, S, Bates, D, Neri, F, Diegel, M, Giste, E, Dunn, D, Vierstra, J, Hansen, RS, Johnson, AK, Sabo, PJ, Wilken, MS, Reh, Ta, Treuting, PM, Kaul, R, Groudine, M, Bender, Ma, Borenstein, E, and Stamatoyannopoulos, JA (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, **515**:365–370.

[327] Hesselberth, JR, Chen, X, Zhang, Z, Sabo, PJ, Sandstrom, R, Reynolds, AP, Thurman, RE, Neph, S, Kuehn, MS, Noble, WS, Fields, S, and Stamatoyannopoulos, JA (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, **6**:283–289.

[328] Lynch, M and Hagner, K (2015). Evolutionary meandering of intermolecular interactions along the drift barrier. *Proceedings of the National Academy of Sciences of the United States of America*, **112**:E30–E38.

[329] MacArthur, S and Brookfield, JFY (2004). Expected rates and modes of evolution of enhancer sequences. *Molecular Biology and Evolution*, **21**:1064–1073.

[330] Bergström, A, Simpson, JT, Salinas, F, Barré, B, Parts, L, Zia, A, Nguyen Ba, AN, Moses, AM, Louis, EJ, Mustonen, V, Warringer, J, Durbin, R, and Liti, G (2014). A high-definition view of functional genetic variation from natural yeast genomes. *Molecular Biology and Evolution*, **31**:872–888.

[331] MacIsaac, KD, Wang, T, Gordon, DB, Gifford, DK, Stormo, GD, and Fraenkel, E

(2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**:113.

[332] Gompel, N, Prud'homme, B, Wittkopp, PJ, Kassner, Va, and Carroll, SB (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature*, **433**:481–487.

[333] Rister, J, Razzaq, A, Boodram, P, Desai, N, Tsanis, C, Chen, H, Jukam, D, and Desplan, C (2015). Single-base pair differences in a shared motif determine differential *Rhodopsin* expression. *Science*, **350**:1258–1261.

[334] Siggers, T and Gordân, R (2014). Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Research*, **42**:2099–2111.

[335] Li, XY, MacArthur, S, Bourgon, R, Nix, D, Pollard, Da, Iyer, VN, Hechmer, A, Simirenko, L, Stapleton, M, Luengo Hendriks, CL, Hou, CC, Ogawa, N, Inwood, W, Sementchenko, V, Beaton, A, Weiszmann, R, Celniker, SE, Knowles, DW, Gingeras, T, Speed, TP, Eisen, MB, and Biggin, MD (2008). Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLOS Biology*, **6**:e27.

[336] Fisher, WW, Li, JJ, Hammonds, AS, Brown, JB, Pfeiffer, BD, Weiszmann, R, MacArthur, S, Thomas, S, Stamatoyannopoulos, JA, Eisen, MB, Bickel, PJ, Biggin, MD, and Celniker, SE (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, **109**:21330–21335.

[337] Mustonen, V and Lässig, M (2009). From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, **25**:111–119.

[338] Arbiza, L, Gronau, I, Aksoy, Ba, Hubisz, MJ, Gulko, B, Keinan, A, and Siepel, A

(2013). Genome-wide inference of natural selection on human transcription factor binding sites. *Nature Genetics*, **45**:723–729.

[339] Mustonen, V and Lässig, M (2005). Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proceedings of the National Academy of Sciences of the United States of America*, **102**:15936–15941.

[340] Swanson, CI, Schwimmer, DB, and Barolo, S (2011). Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Current Biology*, **21**:1186–1196.

[341] Grönlund, A, Lötstedt, P, and Elf, J (2013). Transcription factor binding kinetics constrain noise suppression via negative feedback. *Nature communications*, **4**:1864.

[342] Ramos, AI and Barolo, S (2013). Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*, **368**:20130018.

[343] Crocker, J, Abe, N, Rinaldi, L, McGregor, AP, Frankel, N, Wang, S, Alsawadi, A, Valenti, P, Plaza, S, Payre, F, Mann, RS, and Stern, DL (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, **160**:191–203.

[344] Grant, CE, Bailey, TL, and Noble, WS (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **27**:1017–1018.

[345] van Helden, J, André, B, and Collado-Vides, J (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, **281**:827–842.

[346] Trapnell, C, Williams, Ba, Pertea, G, Mortazavi, A, Kwan, G, van Baren, MJ, Salzberg, SL, Wold, BJ, and Pachter, L (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**:511–515.

[347] Quinlan, AR and Hall, IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**:841–842.

[348] Franke, J, Klözer, A, de Visser, JAGM, and Krug, J (2011). Evolutionary accessibility of mutational pathways. *PLOS Computational Biology*, **7**:e1002134.

[349] Parker, DS, White, MA, Ramos, AI, Cohen, BA, and Barolo, S (2011). The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Science Signaling*, **4**:ra38.

[350] ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**:57–74.

[351] Sokal, R and Rohlf, F (1995). *Biometry*. Freeman, San Francisco, 3rd edition.

[352] du Plessis, L, Leventhal, GE, and Bonhoeffer, S (2016). How good are statistical models at approximating complex fitness landscapes? *Molecular Biology and Evolution*, **33**:2454–2468.

[353] Breen, MS, Kemena, C, Vlasov, PK, Notredame, C, and Kondrashov, FA (2012). Epistasis as the primary factor in molecular evolution. *Nature*, **490**:535–538.

[354] McCandlish, DM, Rajon, E, Shah, P, Ding, Y, and Plotkin, JB (2013). The role of epistasis in protein evolution. *Nature*, **497**:E1–2; discussion E2–3.

[355] Man, TK and Stormo, GD (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Research*, **29**:2471–2478.

[356] Bulyk, ML, Johnson, PLF, and Church, GM (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, **30**:1255–1261.

[357] Anderson, DW, McKeown, AN, and Thornton, JW (2015). Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife*, **4**:e07864.

[358] Dubertret, B, Liu, S, Ouyang, Q, and Libchaber, A (2001). Dynamics of DNA-protein interaction deduced from in vitro DNA evolution. *Physical Review Letters*, **86**:6022–6025.

[359] von Hippel, PH and Berg, OG (1986). On the specificity of DNA-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **83**:1608–1612.

[360] Berg, OG, Hippel, PH, and von Hippel, PH (1988). Selection of DNA binding sites by regulatory proteins. *Trends in Biochemical Sciences*, **13**:207–211.

[361] Sengupta, AM, Djordjevic, M, and Shraiman, BI (2002). Specificity and robustness in transcription control networks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**:2072–2077.

[362] Gerland, U, Moroz, JD, and Hwa, T (2002). Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, **99**:12015–12020.

[363] Tuğrul, M, Paixão, T, Barton, NH, and Tkačik, G (2015). Dynamics of transcription factor binding site evolution. *PLOS Genetics*, **11**:e1005639.

[364] Buchler, NE, Gerland, U, and Hwa, T (2003). On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences of the United States of America*, **100**:5136–5141.

[365] Raveh-Sadka, T, Levo, M, Shabi, U, Shany, B, Keren, L, Lotan-Pompan, M, Zeevi, D, Sharon, E, Weinberger, A, and Segal, E (2012). Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nature Genetics*, **44**:743–750.

[366] Levo, M, Zalckvar, E, Sharon, E, Machado, ACD, Kalma, Y, Lotam-Pompan, M, Weinberger, A, Yakhini, Z, Rohs, R, and Segal, E (2015). Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research*, **25**:1018–1029.

[367] Gordân, R, Hartemink, AJ, and Bulyk, ML (2009). Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Research*, **19**:2090–2100.

[368] de Vos, MGJ, Dawid, A, Sunderlikova, V, and Tans, SJ (2015). Breaking evolutionary constraint with a tradeoff ratchet. *Proceedings of the National Academy of Sciences of the United States of America*, **112**:14906–14911.

[369] Baker, CR, Tuch, BB, and Johnson, AD (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proceedings of the National Academy of Sciences of the United States of America*, **108**:7493–7498.

[370] Sayou, C, Monniaux, M, Nanao, MH, Moyroud, E, Brockington, SF, Thévenon, E, Chahtane, H, Warthmann, N, Melkonian, M, Zhang, Y, Wong, GKS, Weigel, D, Parcy, F, and Dumas, R (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science*, **343**:645–648.

[371] Maerkl, SJ and Quake, SR (2009). Experimental determination of the evolvability of a transcription factor. *Proceedings of the National Academy of Sciences of the United States of America*, **106**:18650–18655.

[372] Nadimpalli, S, Persikov, AV, and Singh, M (2015). Pervasive variation of transcription factor orthologs contributes to regulatory network evolution. *PLOS Genetics*, **11**:e1005011.

[373] Pigliucci, M (2010). Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. *Proceedings of the Royal Society B*, **365**:557–566.

[374] Greenbury, SF, Johnson, IG, Louis, AA, and Ahnert, SE (2014). A tractable

genotype-phenotype map modelling the self-assembly of protein quaternary structure. *Journal of the Royal Society Interface*, **11**:20140249.

[375] Bornberg-Bauer, E and Chan, H (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences of the United States of America*, **96**:10689–10694.

[376] Hu, T, Payne, JL, Banzhaf, W, and Moore, JH (2012). Evolutionary dynamics on multiple scales: a quantitative analysis of the interplay between genotype, phenotype, and fitness in linear genetic programming. *Genet Program Evol M*, **13**:305–337.

[377] Payne, JL and Wagner, A (2013). Constraint and contingency in multifunctional gene regulatory circuits. *PLOS Computational Biology*, **9**:e1003071.

[378] Ferrada, E (2014). The amino acid alphabet and the architecture of the protein sequence-structure map. I. Binary alphabets. *PLOS Computational Biology*, **10**:e1003946.

[379] Louis, AA (2016). Contingency, convergence and hyper-astronomical numbers in biological evolution. *Studies in History and Philosophy of Biological and Biomedical Sciences*, **58**:107–116.

[380] Guerreiro, I, Nunes, A, Woltering, JM, Casaca, A, Nóvoa, A, Vinagre, T, Hunter, ME, Duboule, D, and Mallo, M (2013). Role of a polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine. *Proceedings of the National Academy of Sciences of the United States of America*, **110**:10682–10686.

[381] Newman, MEJ (2010). *Networks: An Introduction*. Oxford University Press, Oxford.

[382] Watts, DJ and Strogatz, SH (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**:440–442.

[383] Newman, MEJ (2002). Assortative mixing in networks. *Phys Rev Lett*, **89**:208701.

[384] Gastner, MT and Newman, MEJ (2006). Shape and efficiency in spatial distribution networks. *Journal of Statistical Mechanics: Theory and Experiment*, P01015.

[385] Zhang, X, Martin, T, and Newman, M (2015). Identification of core-periphery structure in networks. *Physical Reviews E*, **91**:032803.

[386] Peel, L, Larremore, DB, and Clauset, A (2017). The ground truth about metadata and community detection in networks. *Science Advances*, **3**:e1602548.

[387] Friedlander, T, Prizak, R, Guet, CC, Barton, NH, and Tkačik, G (2016). Intrinsic limits to gene regulation by global crosstalk. *Nature Communications*, **7**:12307.

[388] Waddington, CH (1959). *The Strategy of the Genes*. Macmillan, New York, US.

[389] West-Eberhard, MJ (2003). *Developmental Plasticity and Evolution*. Oxford University Press, Oxford, UK.

[390] Bloom, JD, Labthavikul, ST, Otey, CR, and Arnold, FH (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, **103**:5896–5874.

[391] Hu, S, Wan, J, Su, Y, Song, Q, Zeng, Y, Nguyen, HN, Shin, J, Cox, E, Rho, HS, Woodard, C, Xia, S, Liu, S, Lyu, H, Ming, GL, Wade, H, Song, H, Qian, J, and Zhu, H (2013). DNA methylation presents distinct binding sites for human transcription factors. *eLife*, **2**:e00726.

[392] Isakova, A, Groux, R, Imbeault, M, Rainer, P, Alpern, D, Dainese, R, Ambrosini, G, Trono, D, Bucher, P, and Deplancke, B (2017). SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nature Methods*, **14**:316–322.

[393] Levo, M, Avnit-Sagi, T, Lotan-Pompan, M, Kalma, Y, Weinberger, A, Yakhini, Z, and Segal, E (2017). Systematic investigation of transcription factor activity in the context of chromatin using massively parallel binding and expression assays. *Molecular Cell*, **65**:604–617.

[394] The UniProt Consortium (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, **43**:D204–D212.

[395] Holland, PW, Laskey, KB, and Leinhardt, S (1983). Stochastic blockmodels: First steps. *Social Networks*, **5**:109–137.

[396] Nowicki, K and Snijders, TAB (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, **96**:1077–1087.

[397] Peixoto, TP (2014). Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Reviews E*, **89**:012804.

[398] Finn, RD, Bateman, A, Clements, J, Coggill, P, Eberhardt, RY, Eddy, SR, Heger, A, Hetherington, K, Holm, L, Mistry, J, Sonnhammer, ELL, Tate, J, and Punta, M (2014). Pfam: the protein families database. *Nucleic Acids Research*, **42**:D222–D230.

[399] Cartwright, R (2009). Problems and solutions for estimating indel rates and length distributions. *Molecular Biology and Evolution*, **26**:473–480.

[400] Chen, JQ, Wu, Y, Yang, H, Bergelson, J, Kreitman, M, and Tian, D (2009). Variation in the ratio of nucleotide subsitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution*, **26**:1523–1531.

[401] de Visser, JAGM, Hermisson, J, Wagner, GP, Ancel Meyers, L, Bagheri-Chaichian, H, Blanchard, JL, Chao, L, Cheverud, JM, Elena, SF, Fontana, W, Gibson, G, Hansen, TF, Krakauer, D, Lewontin, RC, Ofria, C, Rice, SH, von Dassow, G, Wagner, A, and Whitlock, MC (2003). Perspective: Evolution and detection of genetic robustness. *Evolution*, **57**:1959–1972.

[402] Masel, J and Siegal, ML (2009). Robustness: mechanisms and consequences. *Trends in Genetics*, **25**:395–403.

[403] Fares, MA (2015). The origins of mutational robustness. *Trends in Genetics*, **31**:373–381.

[404] Young, JC, Agashe, VR, Siegers, K, and Hartl, FU (2004). Pathways of chaperone-mediated protein folding in the cytosol. *Nature Reviews Molecular Cell Biology*, 781–791.

[405] Rudan, M, Schneider, D, Warnecke, T, and Krisko, A (2015). RNA chaperones buffer deleterious mutations in E. coli. *eLife*, **4**:1–16.

[406] Cowen, LE and Lindquist, S (2005). Hsp90 potentiates the rapid evolution of new traits: drug resistance in diverse fungi. *Science*, **309**:2185–2189.

[407] Burga, A, Casanueva, MO, and Lehner, B (2011). Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature*, **480**:250–253.

[408] Lachowiec, J, Lemus, T, Thomas, JH, Murphy, PJM, Nemhauser, JL, and Queitsch, C (2013). The Protein Chaperone HSP90 Can Facilitate the Divergence of Gene Duplicates. *Genetics*, **193**:1269–1277.

[409] Pechmann, S and Frydman, J (2014). Interplay between Chaperones and Protein Disorder Promotes the Evolution of Protein Networks. *PLOS Computational Biology*, **10**:e1003674.

[410] Bershtein, S, Mu, W, Serohijos, AWR, Zhou, J, and Shakhnovich, EI (2013). Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Molecular Cell*, **49**:133–144.

[411] Wyganowski, KT, Kaltenbach, M, and Tokuriki, N (2013). GroEL/ES buffering and compensatory mutations promote protein evolution by stabilizing folding intermediates. *Journal of Molecular Biology*, **425**:3403–3414.

[412] Warnecke, T and Hurst, LD (2010). GroEL dependency affects codon usage–support for a critical role of misfolding in gene evolution. *Molecular Systems Biology*, **6**:340.

[413] Pedersen, KS, Kristensen, TN, and Loeschcke, V (2005). Effects of inbreeding and rate of inbreeding in *Drosophila melanogaster*-Hsp70 expression and fitness. *Journal of Evolutionary Biology*, **18**:756–762.

[414] Powers, ET and Balch, WE (2013). Diversity in the origins of proteostasis networks - a driver for protein function in evolution. *Nature Reviews Molecular Cell Biology*, **14**:237–248.

[415] Warnecke, T (2012). Loss of the DnaK-DnaJ-GrpE chaperone system among the Aquificales. *Molecular Biology and Evolution*, **29**:3485–3495.

[416] Eyre-Walker, A and Keightley, PD (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 610–618.

[417] Nishihara, K, Kanemori, M, Kitagawa, M, Yanagi, H, and Yura, T (1998). Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in *Escherichia coli*. *Applied and Environmental Microbiology*, **64**:1694–1699.

[418] Kim, YE, Hipp, MS, Bracher, A, Hayer-Hartl, M, and Hartl, FU (2013). Molecular chaperone functions in protein folding and proteostasis. *Annual Review of Biochemistry*, **82**:323–355.

[419] Turrientes, MC, Baquero, F, Levin, BR, Martínez, JL, Ripoll, A, González-Alba, JM, Tobes, R, Manrique, M, Baquero, MR, Rodríguez-Domínguez, MJ, Cantón, R, and Galán, JC (2013). Normal mutation rate variants arise in a Mutator (Mut S) *Escherichia coli* population. *PLOS ONE*, **8**:e72963.

[420] Bradford, MM (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, **72**:248–254.

[421] Schneider, CA, Rasband, WS, and Eliceiri, KW (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, **9**:671–675.

[422] Deatherage, DE and Barrick, JE (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. *Methods in Molecular Biology*, **1151**:165–188.

[423] Langmead, B and Salzberg, SL (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**:357–359.

[424] Tatusov, RL, Koonin, E, and Lipman, DJ (1997). A genomic perspective on protein families. *Science*, **278**:631–637.

[425] Rice, P, Longden, I, and Bleasby, A (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**:276–277.

[426] Suyama, M, Torrents, D, and Bork, P (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**:W609–W612.

[427] Yang, Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**:1586–1591.

[428] Felsenstein, J (2005). PHYLIP (Phylogeny Inference Package) version 3.6.

[429] Jones, DT, Taylor, WR, and Thornton, JM (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, **8**:275–282.

[430] Rajagopala, SV, Sikorski, P, Kumar, A, Mosca, R, Vlasblom, J, Arnold, R, Franca-Koh, J, Pakala, SB, Phanse, S, Ceol, A, Häuser, R, Siszler, G, Wuchty, S, Emili, A, Babu, M, Aloy, P, Pieper, R, and Uetz, P (2014). The binary protein-protein interaction landscape of *Escherichia coli*. *Nature Biotechnology*, **32**:285–290.

[431] Baba, T, Ara, T, Hasegawa, M, Takai, Y, Okumura, Y, Baba, M, Datsenko, Ka, Tomita, M, Wanner, BL, and Mori, H (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, **2**:2006.0008.

[432] Chen, X and Zhang, J (2013). No gene-specific optimization of mutation rate in *Escherichia coli*. *Molecular Biology and Evolution*, **30**:1559–1562.

[433] R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[434] Wall, DP, Hirsh, AE, Fraser, HB, Kumm, J, Giaever, G, Eisen, MB, and Feldman, MW (2005). Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **102**:5483–5488.

[435] Drummond, DA, Raval, A, and Wilke, CO (2006). A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, **23**:327–337.

[436] Mills, DR, Peterson, RL, and Spiegelman, S (1967). An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proceedings of the National Academy of Sciences of the United States of America*, **58**:217–24.

[437] Futuyma, DJ (1988). The evolution of ecological specialization. *Annual Review of Ecology and Systematics*, **19**:207–233.

[438] Bennett, AF and Lenski, RE (1993). Evolutionary adaptation to temperature II. Thermal niches of experimental lines of *Escherichia coli*. *Evolution*, **47**:1–12.

[439] Fong, DW, Kane, TC, and Culver, DC (1995). Vestigialization and loss of nonfunctional characters. *Annual Review of Ecology and Systematics*, **26**:249–268.

[440] Cooper, VS and Lenski, RE (2000). The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature*, **407**:736–739.

[441] Cooper, VS (2014). The origins of specialization: insights from bacteria held 25 years in captivity. *PLOS Biology*, **12**:e1001790.

[442] Leiby, N and Marx, CJ (2014). Metabolic Erosion Primarily Through Mutation Accumulation, and Not Tradeoffs, Drives Limited Evolution of Substrate Specificity in *Escherichia coli*. *PLOS Biology*, **12**(2):e1001789.

[443] Moran, NA, McLaughlin, HJ, and Sorek, R (2009). The Dynamics and Time Scale of Ongoing Genomic Erosion in Symbiotic Bacteria. *Science*, **323**:379–382.

[444] Moran, NA and Wernegreen, JJ (2000). Lifestyle evolution in symbiotic bacteria: Insights from genomics. *Trends in Ecology and Evolution*, **15**:321–326.

[445] Lenski, RE, Rose, MR, Simpson, SC, and Tadler, SC (1991). Long-Term Experimental Evolution in *Escherichia coli*. I. Adaptation and Divergence During. *The American Naturalist*, **138**(6):1315–1341.

[446] Sniegowski, PD, Gerrish, PJ, and Lenski, RE (1997). Evolution of high mutation rates in experimental populations of *E. coli*. *Nature*, **387**:703–705.

[447] Takemoto, K, Niwa, T, and Taguchi, H (2011). Difference in the distribution pattern of substrate enzymes in the metabolic network of *Escherichia coli*, according to chaperonin requirement. *BMC Systems Biology*, **5**:98.

[448] Alvarez-Ponce, D, Sabater-Muñoz, B, Toft, C, Ruiz-González, MX, and Fares, MA (2016). Essentiality is a strong determinant of protein rates of evolution during mutation accumulation experiments in *Escherichia coli*. *Genome Biology and Evolution*, **8**:2914–2927.

[449] Sabater-Muñoz, B, Toft, C, Alvarez-Ponce, D, and Fares, MA (2017). Chance and necessity in the genome evolution of endosymbiotic bacteria of insects. *The ISME Journal*, 1–14.

[450] Bochner, BR, Gadzinski, P, and Panomitros, E (2001). Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Research*, **11**:1246–1255.

[451] Bochner, BR (2009). Global phenotypic characterization of bacteria. *FEMS Microbiology Reviews*, **33**:191–205.

[452] Zhou, L, Lei, XH, Bochner, BR, and Wanner, BL (2003). Phenotype microarray analysis of *Escherichia coli* K-12 mutants with deletions of all two-component systems. *Journal of Bacterioly*, **185**:4956–4972.

[453] Pommerenke, C, Müsken, M, Becker, T, Dötsch, A, Klawonn, F, and Haussler, S (2010). Global genotype-phenotype correlations in *Pseudomonas aeruginosa*. *PLOS Pathogens*, **6**:89–90.

[454] Toll-Riera, M, San Millan, A, Wagner, A, and MacLean, RC (2016). The genomic basis of evolutionary innovation in *Pseudomonas aeruginosa*. *PLOS Genetics*, **12**:e1006005.

[455] Loh, KD, Gyaneshwar, P, Markenscoff Papadimitriou, E, Fong, R, Kim, KS, Parales, R, Zhou, Z, Inwood, W, and Kustu, S (2006). A previously undescribed pathway for pyrimidine catabolism. *Proceedings of the National Academy of Sciences*, **103**:5114–5119.

[456] Sabarly, V, Bouvet, O, Glodt, J, Clermont, O, Skurnik, D, Diancourt, L, De Vienne, D, Denamur, E, and Dillmann, C (2011). The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity. *Journal of Evolutionary Biology*, **24**:1559–1571.

[457] Plata, G, Henry, CS, and Vitkup, D (2015). Long-term phenotypic evolution of bacteria. *Nature*, **517**:369–372.

[458] Le Gac, M, Plucain, J, Hindre, T, Lenski, RE, and Schneider, D (2012). Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment

with *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, **109**:9487–9492.

[459] Hug, SM and Gaut, BS (2015). The phenotypic signature of adaptation to thermal stress in *Escherichia coli*. *BMC Evolutionary Biology*, **15**:177.

[460] Keseler, IM, Mackie, A, Peralta-Gil, M, Santos-Zavaleta, A, Gama-Castro, S, Bonavides-Martínez, C, Fulcher, C, Huerta, AM, Kothari, A, Krummenacker, M, Latendresse, M, Muñiz-Rascado, L, Ong, Q, Paley, S, Schröder, I, Shearer, AG, Subhraveti, P, Travers, M, Weerasinghe, D, Weiss, V, Collado-Vides, J, Gunsalus, RP, Paulsen, I, and Karp, PD (2013). EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Research*, **41**:605–612.

[461] Lee, H, Popodi, E, Tang, H, and Foster, PL (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **109**:E2774–83.

[462] Nilsson, AI, Koskiniemi, S, Eriksson, S, Kugelberg, E, Hinton, JCD, and Andersson, DI (2005). Bacterial genome size reduction by experimental evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **102**:12112–12116.

[463] Li, WH, Wu, CI, and Luo, CC (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, **2**:150–174.

[464] Alvarez-Ponce, D (2014). Why proteins evolve at different rates: The determinants of proteins' rates of evolution. In MA Fares, editor, *Natural Selection: Methods and Applications*, CRC Press (Taylor & Francis), 126–178.

[465] Duret, L and Mouchiroud, D (2000). Determinants of substitution rates in mam-

malian genes: Expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution*, **17**:68–74.

[466] Subramanian, S and Kumar, S (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, **168**:373–381.

[467] Liao, BY, Scott, NM, and Zhang, J (2006). Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Molecular Biology and Evolution*, **23**:2072–2080.

[468] Ingvarsson, PK (2007). Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. *Molecular Biology and Evolution*, **24**:836–844.

[469] Kryuchkova, N and Robinson-Rechavi, M (2014). Determinants of protein evolutionary rates in light of ENCODE functional genomics. *BMC Bioinformatics*, **15**.

[470] Hurst, LD and Smith, NGC (1999). Do essential genes evolve slowly? *Current Biology*, **9**:747–750.

[471] Jordan, IK, Rogozin, IB, Wolf, YI, and Koonin, EV (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Research*, **12**:962–968.

[472] Wilson, AC, Carlson, SS, and White, TJ (1977). Biochemical evolution. *Annual Review of Biochemistry*, **46**:573–639.

[473] Salathé, M, Ackermann, M, and Bonhoeffer, S (2006). The effect of multifunctionality on the rate of evolution in yeast. *Molecular Biology and Evolution*, **23**:721–722.

[474] He, X and Zhang, J (2006). Toward a molecular understanding of pleiotropy. *Genetics*, **173**:1885–1891.

[475] Podder, S, Mukhopadhyay, P, and Ghosh, TC (2009). Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. *Gene*, **439**:11–16.

[476] Liao, BY, Weng, MP, and Zhang, J (2010). Impact of extracellularity on the evolutionary rate of mammalian proteins. *Genome Biology and Evolution*, **2**:39–43.

[477] Cork, JM and Purugganan, MD (2004). The evolution of molecular genetic pathways and networks. *BioEssays*, **26**:479–484.

[478] Fraser, HB, Hirsh, AE, Steinmetz, LM, Scharfe, C, and Feldman, MW (2002). Evolutionary rate in the protein interaction network. *Science*, **296**:750–752.

[479] Jordan, IK, Wolf, YI, and Koonin, EV (2003). No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evolutionary Biology*, **3**.

[480] Hahn, MW and Kern, AD (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, **22**:803–806.

[481] Lemos, B, Bettencourt, BR, Meiklejohn, CD, and Hartl, DL (2005). Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Molecular Biology and Evolution*, **22**:1345–1354.

[482] Alvarez-Ponce, D (2012). The relationship between the hierarchical position of proteins in the human signal transduction network and their rate of evolution. *BMC Evolutionary Biology*, **12**:192.

[483] Alvarez-Ponce, D and Fares, MA (2012). Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein-protein interaction network. *Genome Biology and Evolution*, **4**:1263–1274.

[484] Jovelin, R and Phillips, PC (2009). Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biology*, **10**:R35.

[485] Barabási, AL and Oltvai, ZN (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**:101–113.

[486] Bloom, JD and Adami, C (2003). Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evolutionary Biology*, **3**:21.

[487] Batada, NN, Hurst, LD, and Tyers, M (2006). Evolutionary and physiological importance of hub proteins. *PLOS Computational Biology*, **2**:e88.

[488] Lu, C, Zhang, Z, Leach, L, Kearsey, M, and Luo, Z (2007). Impacts of yeast metabolic network structure on enzyme evolution. *Genome Biology*, **8**:407.

[489] Greenberg, AJ, Stockwell, SR, and Clark, AG (2008). Evolutionary constraint and adaptation in the metabolic network of *Drosophila*. *Molecular Biology and Evolution*, **25**:2537–2546.

[490] Hudson, C and Conant, G (2011). Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes. *BMC Evolutionary Biology*, **11**:89.

[491] Montanucci, L, Laayouni, H, Dall'Olio, GM, and Bertranpetit, J (2011). Molecular evolution and network-level analysis of the N-glycosylation metabolic pathway across primates. *Molecular Biology and Evolution*, **28**:813–823.

[492] Colombo, M, Laayouni, H, Invergo, BM, Bertranpetit, J, and Montanucci, L (2014). Metabolic flux is a determinant of the evolutionary rates of enzyme-encoding genes. *Evolution*, **68**:605–613.

[493] Samal, A, Matias Rodrigues, JF, Jost, J, Martin, OC, and Wagner, A (2010). Genotype networks in metabolic reaction spaces. *BMC Systems Biology*, **4**:30.

[494] Hahn, MW, Conant, GC, and Wagner, A (2004). Molecular evolution in large genetic networks: Does connectivity equal constraint? *Journal of Molecular Evolution*, **58**:203–211.

[495] Wang, M, Weiss, M, Simonovic, M, Haertinger, G, Schrimpf, SP, Hengartner, MO, and von Mering, C (2012). PaxDb, a database of protein abundance averages across all three domains of life. *Molecular & Cellular Proteomics*, **11**:492–500.

[496] Harcombe, WR, Delaney, NF, Leiby, N, Klitgord, N, and Marx, CJ (2013). The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. *PLOS Computational Biology*, **9**.

[497] Hosseini, SR, Barve, A, and Wagner, A (2015). Exhaustive analysis of a genotype space comprising 10^{15} central carbon metabolisms reveals an organization conducive to metabolic innovation. *PLOS Computational Biology*, **11**:e1004329.

[498] Edwards, JS and Palsson, BØ (2000). The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America*, **97**:5528–5533.

[499] Edwards, JS, Ibarra, RU, and Palsson, BØ (2001). *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, **19**(2):125–130.

[500] Ibarra, RU, Edwards, JS, and Palsson, BØ (2002). *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature*, **420**:186–189.

[501] Segrè, D, Vitkup, D, and Church, GM (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, **99**:15112–15117.

[502] Fong, SS and Palsson, BØ (2004). Metabolic gene-deletion strains of *Escherichia coli*

evolve to computationally predicted growth phenotypes. *Nature Genetics*, **36**:1056–1058.

[503] Lewis, NE, Hixson, KK, Conrad, TM, Lerman, Ja, Charusanti, P, Polpitiya, AD, Adkins, JN, Schramm, G, Purvine, SO, Lopez-Ferrer, D, Weitz, KK, Eils, R, König, R, Smith, RD, and Palsson, BØ (2010). Omic data from evolvedv *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, **6**:390.

[504] Dowell, R, Ryan, O, Jansen, A, Cheung, D, Agarwala, S, Danford, T, Bernstein, D, Rolfe, A, Heisler, L, Chin, B, Nislow, C, Giaever, G, Phillips, P, Fink, G, Gifford, D, and Boone, C (2010). Genotype to phenotype: A complex problem. *Science*, **328**:469.

[505] Edwards, JS and Palsson, BØ (1999). Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *Cell Biology and Metabolism*, **274**:17410–17416.

[506] Edwards, JS and Palsson, BØ (2000). Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnology Progress*, **16**:927–939.

[507] Stern, DL and Orgogozo, V (2008). The loci of evolution: How predictable is genetic evolution? *Evolution*, **62**:2155–2177.

[508] Baatz, M and Wagner, GP (1997). Adaptive inertia caused by hidden pleiotropic effects. *Theoretical Population Biology*, **51**:49–66.

[509] Orr, HA (2000). Adaptation and the cost of complexity. *Evolution*, **54**:13–20.

[510] Otto, SP (2004). Two steps forward, one step back: the pleiotropic effects of favoured alleles. *Proceedings of the Royal Society B*, **271**:705–714.

[511] Cooper, TF, Ostrowski, EA, and Travisano, M (2007). A negative relationship between mutation pleiotropy and fitness effect in yeast. *Evolution*, **61**:1495–1499.

[512] Ochman, H and Jones, IB (2000). Evolutionary Genomics of full genome content in *Escherichia coli*. *The EMBO Journal*, **19**:6637–6643.

[513] Waxman and Peck (1998). Pleiotropy and the preservation of perfection. *Science*, **279**:1210–1213.

[514] Toll-Riera, M, Bostick, D, Albà, MM, and Plotkin, JB (2012). Structure and age jointly influence rates of protein evolution. *PLOS Computational Biology*, **8**.

[515] Montañez, R, Medina, MA, Solé, RV, and Rodríguez-Caso, C (2010). When metabolism meets topology: Reconciling metabolite and reaction networks. *BioEssays*, **32**:246–256.

[516] Freeman, LC (1977). A set of measures of centrality based on betweenness. *Sociometry*, **40**:35.

[517] Schellenberger, J, Que, R, Fleming, RMT, Thiele, I, Orth, JD, Feist, AM, Zielinski, DC, Bordbar, A, Lewis, NE, Rahmanian, S, Kang, J, Hyduke, DR, and Palsson, BØ (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature Protocols*, **6**:1290–1307.

[518] Beard, DA, Liang, SC, and Qian, H (2002). Energy balance for analysis of complex metabolic networks. *Biophysical Journal*, **83**:79–86.

[519] Do, CB, Mahabhashyam, MSP, Brudno, M, and Batzoglou, S (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, **15**:330–340.